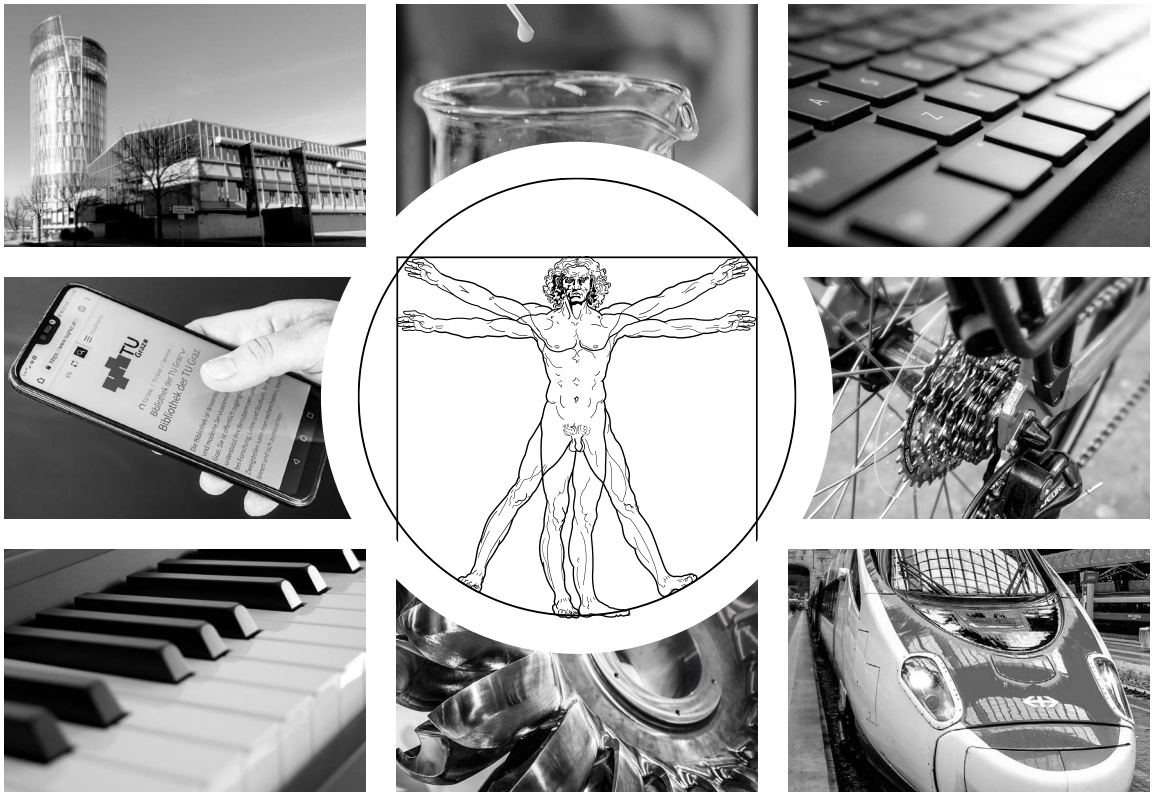
SCIENCE, TECHNOLOGY AND SOCIETY



Günter Getzinger | Michaela Jahrbacher | Roman Prunč (eds.)

Conference Proceedings of the 23rd STS Conference Graz 2025

Critical Issues in Science, Technology and Society Studies

5 – 7 May 2025

TU Graz

Science, Technology and Society

Günter Getzinger, Michaela Jahrbacher, Roman Prunč (eds.)

Conference Proceedings of the

23rd STS Conference Graz 2025

Critical Issues in Science, Technology, and Society Studies

May 5th – 7th 2025

Editors:	Günter Getzinger, Michaela Jahrbacher, Roman Prunč
Layout:	Roman Prunč
Cover Design:	Verlag der Technischen Universität Graz
Cover Pictures:	Bernhard Wieser (Science Tower)
	Franz Georg Pikl (Turbine)
	Dietmar Herbst (Bicycle; Smartphone)
	Helmut Lunghammer (Chemistry tools)
	Martin Smoliner (Train)
	Stefan Schleich (Piano)
	MysticaLink / Shutterstock.com (The Vitruvian Man)
	Shaba.One / Shutterstock.com (Keyboard)

2026 Verlag der Technischen Universität Graz

www.tugraz-verlag.at

E-Book

ISBN 978-3-99161-062-5

DOI 10.3217/978-3-99161-062-5

Preface

Critical Issues in Science, Technology and Society Studies

Conference Proceedings of the STS Conference Graz 2025, May 5th – 7th

The annual STS Conference Graz provides a space for scholars from all parts of the world to present and discuss their research with peers. In their papers, the conference participants address the complex ways in which science, technology and society coevolve and mutually shape one another. Without exception, the participants of the conference aim to provide a better understanding of the world(s) in which we live. This includes the assessment of emerging technologies, the scrutiny of ethical, legal and social aspects of contemporary scientific practices as well as the transition to environmentally friendly and socially desirable techno-scientific futures.

This volume of proceeding documents is part of the work that has been presented at the 23rd STS Conference in Graz in 2025. It presents the wealth of ideas discussed at this occasion and fosters collaboration. The STS Conference Graz is the joint annual conference of the Science, Technology and Society (STS) Unit at Graz University of Technology, the Interdisciplinary Research Centre for Technology, Work and Culture (IFZ) and the Institute for Advanced Studies on Science, Technology and Society (IASSTS).

Find the Book of Abstracts at the DOI 10.3217/978-3-99161-054-0

Contents

Digital city and disaster twins – towards a critical understanding of cyber-physical governance

Kathrin Braun

University of Stuttgart, Germany

DOI 10.3217/978-3-99161-062-5-001, CC BY 4.0

Abstract. Digital city twins are a rapidly spreading phenomenon in digital urbanism. A particular variant are digital city twins for purposes of disaster management. The paper explores the vision of digital city twins for disaster management from a critical, STS-informed perspective. The concept of digital twin has been criticised for implying a realist epistemology and the idea that digital twins could provide an accurate representation of physical reality. This paper highlights an emerging literature that understands digital twins not as realist representations but as self-regulating cyber-physical systems capable of immediate analysis and at least partly automated self-regulating intervention based on real-time bidirectional flows of data and information between the physical and the virtual object. It argues that the vision of self-regulating digital disaster twins requires a different type of critique than the critique of epistemological realism, one that focuses on the question what actually is the system that is supposed to regulate and preserve itself. In this respect, critical analysis can derive key questions and insights from critical studies on disaster and emergency management. Drawing especially on the work of David Keen, the paper argues that the vision of a self-regulating digital disaster twin could further depoliticize disaster management, obscure the relations of power and inequality underlying the causes for and the management of disasters and, in the worst case, lead to an automated biopolitics of disposability. It concludes with the question what kind of political debate is needed to prevent this scenario.

1. Introduction

Digital twins increasingly populate our world; today, there are digital twins, or at least digital twin concepts or projects, for aircrafts, buildings, power plants, cities, newborns, tomato farms, oceans, and more. In 2021, the European Union even launched the initiative of building a digital twin of the earth[1]. The universality of the concept is striking and perhaps unprecedented, in particular when we consider that digital twins aspire to be not just a metaphor but an applicable technological approach.

In this paper, I will focus on digital twins for cities, particularly digital city twins for purposes of disaster management, and explore what a critical, STS-informed understanding of this concept might mean. The paper is based on the relevant literature and takes a theoretical perspective. I venture the thesis that what differentiates the digital twin concept from 3D models and visualisations, digital maps, simulations or dashboards, is a cybernetic vision of systemic cyber-physical self-regulation underlying it and that this poses distinct and novel questions for critical STS analysis. What questions need to be asked and which kind of issues and challenges need to be addressed if we take the cybernetic understanding of digital city twins seriously, in particular concerning the purpose of disaster management?

In the following, I will first revisit the concept of digital twin and argue that next to the prevailing (mis)understanding of digital twin as an accurate representation of reality there is also an emerging literature that conceptualises digital twins within a cybernetic framework as cyber-physical systems capable of data-based self-regulation. Next, I introduce the concept of digital city twins and argue that what distinguished them conceptually from other digital models, simulations and technologies is the capacity of 'now-casting', that is the capacity of immediate analysis and response to current conditions based on real-time bidirectional flows of data and information between the physical and the virtual object. I argue that the vision of digital city twins as self-regulation cyber-physical systems merits critical STS-informed analysis and assessment, which becomes particularly clear when we consider digital city twins for disaster and emergency management. The remainder of the paper is devoted to my argument that critical reflexions on digital disaster twins can derive key questions and insights from critical studies on disaster and emergency management. Against the backdrop of David Keen's (2023) critique of disaster and emergency management, the paper argues that the vision of a self-regulating digital disaster twin, if it became true, could further depoliticize disaster management and the underlying relations of power and inequality and, in the worst case, lead to an automation of what Henry Giroux termed the biopolitics of disposability. The

[1] For a critical analysis see Rothe, Delf. 2024. 'When the World Is an Object: On the Governmental Promise of a Digital Twin Earth.' *International Political Sociology* 18(olae022)..

paper concludes with the question what kind of political debate is needed to prevent this scenario.

2 Genealogy and Conceptual Ambiguity

The origins of the digital twin concept have been traced back as far as the 1970s (Lagap and Ghaffarian 2024, Mylonas et al. 2021), but it is only in the past few years that it has given rise to an exponentially growing area of research, a rapidly evolving field of technology applications, and a rapidly evolving market. Historical accounts of its evolution vary, as do definitions. According to one account, the idea of twinning in technology research and development originates in the NASA's Apollo program in 1970, which involved the creation of two identical spacecrafts one of which, named the ‘twin’, stayed on the ground and was used to provide assistance to the crew in space in critical situations (Lagap and Ghaffarian 2024). According to another account, the idea of what was later termed digital twin was first introduced by Michael Grieves in 2002 (Grieves 2002, Grieves 2024) as a model for Product Lifecycle Management. However, the term as such was not used until 2010 by John Vickers of NASA (Grieves 2024) and made more popular in the following years by Grieves and Vickers (Grieves 2014, Grieves and Vickers 2017). Since the mid-2010s, the literature on digital twins is growing exponentially (Ariyachandra and Wedawatta 2023, Cheng, Hou and CHeng 2023, Mohiuddin Eumi 2024).

In technological terms, there is some convergence in the literature saying that a digital twin comprises three basic components:

a) a wide variety of hardware components for collecting data such as IoT sensors, 5G networks, energy and transit data networks, but also official statistics, and social media. In the case of digital city twins, these may be data about the terrain, the built environment, mobility, energy consumption, temperature, and more, up to virus infections and other;
b) a platform for storing, integrating and processing data;
c) the analytical capacity for performing calculations, simulations, scenarios.

Despite the hype around digital twins, or perhaps because of it, there is still a considerable amount of terminological confusion, compounded by a ‘growing eagerness to label everything as a digital twin’ (Martinescu 2023). When doing a rough search on Google, you will quickly come across phrases such as ‘a virtual representation of an object or system designed to reflect a physical object accurately’ (IBM), ‘a virtual replica of a physical object, person, or process’ (MacKinsey & Company), ‘digital images of physical objects’ (Fraunhofer IKS) and the like. Such phrases show an essentially dualist, Cartesian understanding of digital twins (Kitchin and Dawkins 2024), with a human subject on the one hand and a physical world on the other, and the digital twin as an

instrument for seeing and representing an object in the physical world. Within a Cartesian framework, the subject and their intentions and will to knowledge precedes the instrument; the subject uses the instrument to generate knowledge about the physical world which then allows them to form the physical world according to their preceding goals and intentions. This dualist, Cartesian conception of digital twin has been criticised from an STS-perspective for obscuring the performative nature of digital twin and reinforcing an inherent realist epistemology, casting the model as an accurate, comprehensive and objective representation of reality that, as such, would inform better, value-neutral interventions. It has also been criticised for reducing complex, local and contextual relationships to matters of efficiency and control, and for a lack of a human dimension and the absence of people designing, creating and using it (Charitonidou 2022, Kitchin and Dawkins 2024, Korenhof, Blok and Kloppenburg 2021, Korenhof, Giesbers and Sanderse 2023). This line of criticism builds on a longstanding literature in STS, critical geography, critical data studies and related fields that has extensively demonstrated that data as well as numbers, figures, statistics and models are never neutral and objective but always in one way or other value-laden and performative and this criticism certainly applies to digital twins too. However, I want to draw attention to the fact, that next to the (mis)understanding of digital twin as accurate representation of reality, there is also an emerging literature that conceptualises digital twin within a non-Cartesian, systems theory or cybernetic framework. The question I want to discuss is whether a cybernetically inspired conception of digital twins, especially digital city twins, raises novel and different questions and grounds for criticism than the Cartesian conception.

3 Cybernetic Visions

A cybernetic conception of digital twin is advocated by Tomko and Winter (Tomko and Winter 2019) who have suggested to abandon the language of replica, representation, mirror and twin altogether. They argue that, rather than the metaphor of a mirror, the metaphor of a brain in a living organism is more adequate for capturing the specific relation between the virtual and the physical object within a digital twin. Like a brain, ‘the digital counterpart is coupled with the physical realm (cyber–physical) by a nerve system of sensors, actuators, and (information-processing) communication lines’ (Tomko and Winter 2019, 398). For them, the coupling between the virtual and the physical counterpart also includes people communicating and interacting with the physical object as well as with its virtual counterpart. Therefore, they suggest, it would be more appropriate to speak of cyber–physical–social eco-systems than of digital twins.

The idea of creating cyber-physical systems capable of self-steering and self-organisation has a long tradition in cybernetic thinking. Andrew Pickering argues that, at least for the British tradition of cybernetics, the brain as an embodied organ in a living

organism that is constantly adapting to a dynamic environment was the central and distinctive object of interest (Pickering 2010). Within a cybernetic framework, digital twins are not representations of reality but self-organising cyber-physical systems, or essential components of such systems (Casadei et al. 2022, Rocha and Barata 2021, Sepasgozar 2021). Casadei et al. for instance present a vision whereby collectives of digital twins allow cyber-physical systems to facilitate self-organisation processes and 'autonomously organise as a whole towards global goals' (Casadei et al. 2022).

In many respects, cybernetic approaches to digital twins draw on the vision of autonomic computing presented by Kephart and Chess (2003) some twenty years ago. They have argued that existing computing systems at the time had reached a level of complexity that human system integrators would soon be unable to manage. The only option to reconcile ever increasing complexity and enduring innovation would be autonomic computing which they define as 'computing systems that can manage themselves given high-level objectives from administrators.' (Kephart and Chess 2003, 41) The essence of autonomic computing, as they envisioned it, was self-management. Self-management, in turn, comprises a number of further capabilities such as self-configuration, self-optimisation, self-healing and self-protection. Casadei et al. speak of self*- capabilities. Being connected through some form of interaction feedback to the physical object, whether through automated or human decision-making or a combination of both, the virtual twin actively intervenes in the physical object, steering, controlling and optimising it.

Cybernetic conceptions of digital twins have been promoted mainly in the area of industrial processing but recently also in the area of cities, urban planning and urban infrastructures (Liu and Tian 2023). What would it mean to create a digital city twin that functions as a self-regulating, self-organising, self-optimising cyber-physical system? What would be the implications from a critical STS perspective? What questions need to be asked and which kind of issues and challenges need to be addressed if we take the cybernetic approach to digital twins seriously? A self-regulating system is not adequately understood as an instrument or a tool used by a preceding, external subject with preceding ideas and intentions. It is not so much characterised by hierarchically imposed, centralised control and preselected values and priorities than by principles of self-regulation, -organisation, and -optimisation. Therefore, I would argue, the critical question is not so much 'who is behind the steering wheel' (Korenhof, Blok and Kloppenburg 2021, 1766) but 'What is the system?'. In a self-regulating cyber-physical system, there is no one steering wheel and no one driver behind it and not necessarily a predetermined goal either, at least not a detailed one. The whole point of a self-organising system is that it organises and maintains itself by flexibly adapting to a changing environment in ways that cannot be known and predetermined in advance. Therefore, I am not sure that a cybernetically designed city twin would actually be geared at a certain predefined desired state, ideal or substantial goal that can be 'read' from it. If a digital twin actually constitutes

a self-regulating cyber-physical system, it should be capable of defining what constitutes the optimum state under the respective conditions at a given point in time.

The critical question therefore might rather be: What is the system? What is the meaning of systemic self-regulation, self-improvement and self-adaptation? When and why is it considered successful? At which costs? Would there be any room for political deliberation, contest and conflict within such a system and if so, where and when? In the following, I will start to explore these questions with regard to digital city twins with a special focus on digital city twins for disaster management.

4 Twinning the City

Digital city twins, also referred to as digital urban twins, virtual city twins, smart city digital twins, urban scale digital twins and other, are a burgeoning phenomenon in digital urbanism. The digital twin concept has been promoted in the context of smart cities and urban management since around 2017 (Ferré-Bigorra, Casals and Gangolells 2022, WEF 2022). Over the past few years, the literature on digital city twins has grown exponentially (El-Agamy et al. 2024), as well as digital twin projects of cities, regions, or urban infrastructures around the world. At present, most digital city twin projects are still in the concept and planning phases (WEF 2023a) and there is reason to assume that many of these rather refer to a 3D model of the city, a digital map or a data platform without a bidirectional flow of data and information (Ferré-Bigorra, Casals and Gangolells 2022).

There is no standard definition of a digital city twin. Often, digital city twins are defined teleologically; what they are is defined by what they should be or should do, suggesting that they not only can but will and do fulfil critical functions, meet challenges, provide solutions, achieve improvements. For instance, the World Economic Forum defines a digital city twin as ‘a virtual replica of a physical city that enables simulation, monitoring, and control of complex urban scenarios, enhancing efficiency and sustainability.’ (WEF 2023b) Digital twins, so a recent review, ‘enable more informed decision-making and optimize planning, operations, finance, and strategy… help reduce carbon emissions and expedite significant projects… enable the simulation of plans before implementation, allowing for the anticipation of potential challenges’ (El-Agamy et al. 2024, 16), they enhance traffic and mobility management as well as public health and safety, reduce energy consumption, optimise supply chain management, stimulate citizen involvement and more. In short, the WEF proclaims: ‘They have the potential to transform cities into more intelligent entities, leading to high-quality urban development and sustainable growth.’ (WEF 2023a, 4)

In general, the digital city twin is conceived as a tool for urban planning and management that allows for monitoring processes and performance and running virtual what-if

scenarios, and thereby provides actionable information for decision-makers in urban planning, government and management as well as in the private sector.

There are broader and more narrow, demanding concepts of digital city twins. For the latter, a distinct and constitutive feature of digital twins, as compared to other models, simulations and technologies, is the capacity of bidirectional real-time data flows: ‘The big difference from traditional simulations is that digital twins use real-time data for their modelling.’ (Willige 2022) Real-time data flows turn the virtual replica from a static to a dynamic digital object that allows for ‘nowcasting’ (ARC 2024, 1), meaning immediate analysis and response to current conditions. Digital city twins in this sense are complex cyber-physical systems where the physical and the virtual object interact with each other in both directions (Deren, Wenbo and Zhengfeng 2021, 2). They do not only allow for the generation of immersive 3D visualisations, what-if scenarios and simulations but also enable the virtual object to intervene into the physical one by exercising actions of control, adjustment or improvement. While generally, digital city twins are defined by their capability of facilitating and improving decision-making, more demanding concepts require a digital twin to be capable of performing automated decision-making.

At present, we can assume that very few existing projects labelled digital city twin, urban digital twin, smart city digital twin and the like, are based on automated bi-directional, real-time flows of data and information between the physical and the virtual object and allow for automated control, decision-making and intervention (El-Agamy et al. 2024, 30). Nevertheless, it is worth taking the idea seriously, since research and development may push existing digital city twins further in this direction. The need for critical reflexion and discussion, I believe, becomes more obvious when we look at one particular variant of digital city twins, namely digital city twins for disaster and emergency management.

5 Twinning Disaster?

Over the past few years, a series of related concepts and paradigms have been proposed such as Digital Twin Smart City for Disaster Risk Management (Ariyachandra and Wedawatta 2023), digital post-disaster risk management twinning (DPRMT) platform (Lagap and Ghaffarian 2024), Disaster City Digital Twin (DCDT) (Fan et al. 2021), or Smart City Digital Twins (SCDT) for disaster management (Ford and Wolf 2020), Digital Twin for Emergency Management of Civil Infrastructure (EMCI) (Cheng, Hou and CHeng 2023) and other. These propositions commonly start from a tableau of problems and challenges that include a growing world population, the global trend towards ongoing urbanisation, ever complex urban infrastructures, and increasing risks of natural and/or man-made disasters and catastrophes from earthquakes, landslides, climate change and extreme weather events such as floods, draughts and heatwaves, hurricanes, up to terrorist attacks, pandemics and more. Digital twin technology is promoted as a promising

approach to improving the resilience of cities to disasters and emergencies by improving the capacity for disaster preparedness as well as disaster management and mitigation and post-disaster recovery. In short, digital disaster twins are being promoted as an appropriate approach to deal with the increasing complexity of cities, urban regions and infrastructures on the one hand and the increasingly probable, but incalculable, unpredictable and unprecedented occurrence of future disasters and emergencies on the other.

Digital twins for disaster management are basically conceived as decision support systems that employ simulations and scenarios to 'find the best decision for the best outcome' (Doğan, Şahin and Karaarslan 2021, 27). In the pre-disaster situation, digital twins, so the idea, can serve to make predictions about possible future disasters and the damage they may cause, develop emergency response plans and enhance disaster preparedness; in the actual event, they can provide detailed, real-time data about the situation on the ground, generate what-if scenarios to analyse and assess the outcomes of possible responses before implementing them in the real world, inform resource allocation and coordinate rescue efforts more effectively. When it comes to the expected benefits of digital twins in disaster management, resilience is a recurring theme. Digital twins are promoted as an effective approach to improving the resilience of cities and communities to natural disasters, enhancing climate resilience, improving urban resilience, building resilient smart cities, and the like. Overall, we can note a pervasive discursive articulation of digital city twin, disaster, emergency, decision-making, efficient management, and resilience, however without resilience being explained in more detail.

Like digital twins in general, digital disaster twins are composed of data-collecting hardware components, data management components like data platforms, and software components for data analytics and generating simulations and scenarios. Data can be collected by satellites, drones, sensors and cameras installed in buildings, infrastructures, vehicles and more, smartphones (e.g. GPS coordinates), all kind of Internet of Things (IoT) devices, but also from existing data repositories such as Building Information Modeling (BIM), census records, official statistics and other. In addition, peculiar methods of data collection in disaster twin concepts are social sensing and mobile crowd sensing. At this point, the criticism that 'digital twin approaches have been largely ignorant of people and what relates to them' (Charitonidou 2022, 242) requires qualification. These models and concepts actually do assign a key role to people, however it is important to examine which role and what for. Crowdsourcing, here, means harnessing a diverse group of people to collect first-hand, real-time disaster-related information and report it to an online platform. In addition, 'crowdsourcing for data processing refers to approaches in which people implement human-easy and computer-difficult tasks (e.g. labeling images, adding coordinates, tagging reports with categories, etc.) to generate structured, high quality, interpreted data for decision-making or machine learning' (Fan et al. 2021, 4). Crowdsourcing for collecting or processing data, thus, is

based on people's voluntary participation. This is not necessarily the case with social sensing.

Social sensing is a data collection method used to extract information from social media platforms such as Twitter and Facebook with the help of Natural Language Processing (NLP) (Lagap and Ghaffarian 2024). Social sensing has the potential, so the suggestions, to provide real-time data that may serve to indicate the onset of a disaster as well as the location, scale and severity of the damage that has occurred and thereby enhance effective decision-making. It is also mentioned as a method to acquire information about people's behaviour, interactions, movements and emotions in a disaster situation, under the premise to enhance effective decision-making (Fan et al. 2021, Lagap and Ghaffarian 2024). Smartphone users can, for instance, take pictures of damaged buildings or infrastructure and share these, together with the related geolocation data, with the authorities. In addition, tapping into people's smartphones may serve to track people's movements, both in advance, for purposes of generating plausible scenarios, and in real-time for knowing and steering actual movements in disaster situations. Thereby, so the idea, smartphone and social media users can provide data on how people responded to a situation of crisis or disaster and thereby contribute to effective decision-making – whether voluntarily or involuntarily.

But what is `effective´ decision-making in this context? What makes decisions more effective, in which respect and in whose perspective? And what does effective decision-making mean when decisions are not just informed but automated and executed by a cyber-physical system called digital twin?

It is difficult to find more concrete descriptions of how exactly a digital twin could improve the effectiveness and efficiency of disaster management. We learn that automated data flows will inform better, faster and more efficient decisions about the size and composition of rescue teams, equipment, mode of transportation and the like. One more concrete example are flood gauges with automated reporting. But decision-making may also target people and their way of acting: 'To avoid chaos, the analyses' conclusions can be utilised to guide people's movements following the accident.' (Ariyachandra and Wedawatta 2023)

6 Subjects of Twins

At present, it is not clear, whether and to which extent there are already digital disaster twins that enable automated bidirectional data-based interactions between the physical and the virtual object. Only very few applications of digital twin technology in disaster management in the literature do describe the 'V2P [virtual object to physical object] twinning process consistently' in more detail (Zio and Miqueles 2024, 8). As yet, most if not all conceptions for digital disaster twins seem to aim at better, more efficient decisions

of human operators, not automated decisions by the cyber-physical system. In case of the latter, automated decision-making could involve, for instance, the automated closing or opening of flood gates in response to certain data, but it could also mean to direct people out of or into certain spatial areas.

The difficulty in finding more concrete use case descriptions, however, is in fact related to the nature of the matter: After all, the idea and the expected benefit of digital disaster twins is not least that they do *not* rely on pre-defined contingency plans, with pre-defined goals, targets, rules and measures for a pre-defined set of situations. Instead, the advantage of digital disaster twins, so the idea, is precisely that they allow for efficient, flexible, and immediate responses to an unlimited scope of possible events of disaster and emergency, including unprecedented ones for which there is no blueprint yet. Digital twins, in other words, promise to offer an approach that is not geared at concrete scenarios but is sufficiently open, flexible and, in a sense, abstract or universal to deal with situations of disaster, crisis and emergency that cannot be concretised in advance. Ideally, a digital disaster twin is a 'universal tool' (Charitonidou 2022, 243) for dealing with unspecified disasters and emergencies.

In my view, it is important to see that the idea of a 'universal' digital disaster twin - one that will immediately and efficiently 'find the best decision for the best outcome' in any kind of situation – corresponds to a dynamic, self-regulating, self-learning, self-updating cyber-physical system based on continuous, automated, bidirectional data-flows between the physical and the virtual object, where the virtual counterpart does not merely *in*form but *per*forms decision-making. The problem with a 'universal disaster twin', I would contend, is not so much that it would be based on standardised patterns, norms and procedures that are ignorant of local people and local specifics, of people's needs and feelings, the diversity among the population, the socio-economic and demographic features of the community and the like. The universality of the 'ideal' digital disaster twin stems from its capacity to respond to concrete local situations on the basis of real-time data; it is not just executing a set of predefined, in-built standards, rules, norms, and procedures. The dynamic disaster twin does not gloss over people and local specifics, but on the contrary seeks to collect, process and integrate as much data as possible from and about as many concrete local people and specifics as possible. This certainly includes surveillance systems, where people are subject to data collection without consenting or even without registering, but also forms of active engagement where people are encouraged and enrolled to collect data and contribute to improving data quality by, for instance, labelling, categorising, and reporting errors. In short, the digital disaster twin does not ignore people but involves mechanisms of subjecting them to surveillance and control and activating them as responsible data workers in situations of disaster and emergency, thus mechanisms of subjectification in the double sense explicated by Foucault.

7 Sovereign Twin?

There are certainly good reasons to expect that disasters will become more frequent, more complex, more devastating and less manageable in the near future, due to climate change, extreme weather events and environmental destruction, but also due to failed states, corrupt governments, dysfunctional institutions, erosion of international cooperation and more. Hence, there is certainly a strong case for improving disaster prevention, preparedness, mitigation and recovery systems and it is quite conceivable that digital disaster twins hold the potential to make disaster management operations faster, more effective and more efficient. Yet, the question is: what means effective and efficient? In relation to what? At what costs? What actually means 'management'? What counts as benefit, what not? What counts as a disaster, what not? What could be down-sides, social implications, or new risks involved? And where is the place to debate these questions? These are political questions to which there are no merely technological answers. In case of digital disaster twins, I would argue, they should be discussed against the insights of critical disaster and emergency studies. In order to assess the potential of digital disaster twins to improve disaster management, I would hold, we need to critically explore the meaning, ambivalences, problems and pitfalls of disaster and emergency management in the first place in order to be able to reflect on the question whether and if so when and how digital twins could rather resolve or rather exacerbate them. It is not doable within the scope of this paper to provide a comprehensive review of critical disaster and emergency studies. To make a start, I will draw on David Keen's When Disasters Come Home' (Keen 2023). Keen argues that disasters and emergencies do not just happen but are often actively created and exploited or sustained by powerful forces. Hence, disaster and emergency relief will remain patchy and largely futile until the underlying causes for creating or sustaining disasters are not being addressed.

Keen raises a number of points that are of utmost importance for a critical reflection on digital disaster twins: First: What counts as disaster or emergency? Keen points out that the Oxford Dictionary of English and the American Merriam-Webster dictionary both define a disaster as a 'sudden' event that causes great damage, loss or destruction (Keen 2023, 2). Likewise, we could add, an emergency is commonly defined by terms such as 'unforeseen', 'urgent', 'suddenly', 'unexpectedly'[2]. For one thing, this understanding of disaster and emergency excludes situations that also cause great damage, destruction and loss of life but do so over an extended period of time. Also, in the aftermath of

[2] The Cambridge Academic Content Dictionary, for instance, defines emergency as 'a dangerous or serious situation, such as an accident, that happens suddenly or unexpectedly and needs immediate action" (quoted at https://dictionary.cambridge.org/dictionary/english/emergency, last access 26 March 2025) . The Merriam-Webster defines emergency as 'an unforeseen combination of circumstances or the resulting state that calls for immediate action' (https://www.merriam-webster.com/dictionary/emergency, last access 26 March 2025).

disasters, it often turns out that they had not been unexpected at all, but result from long-term underlying problems such as neglect of domestic infrastructure and warning systems, failure to control compliance with building codes and safety regulations or to maintain contingency plans and store medication, protective gear and equipment. In particular, the long-term neglect and underfunding of public health care systems exacerbate disasters and their consequences - or could actually be considered a disaster in itself. The same is true to escalating social inequality, mass poverty, mass incarceration, and poor social security systems. Hence, many disasters and much suffering and loss of life could be averted if it was not for the lack of political will to address these underlying problems. Furthermore, disasters and emergencies tend to be defined by contrasting them with a situation of normality. Accordingly, the function of disaster management is to restore normality as fast and effectively as possible. But what counts as normality? Which interventions to restore 'normality' provide support and relief for whom? And, 'isn't the process of reconstruction and the process of ‚getting back to normal' actually a rather direct route back to the vulnerabilities and grievances that created the disaster in the first place?' (Keen 2023, 39)

One could go further and examine the top priority of disaster and emergency management: Saving lives or saving 'the system'? Whose lives and which 'system'? Keen reminds us of Hurricane Katrina, where people were trapped without water, food or shelter in the flooded parts of the city, which, not coincidentally, were the parts where poorer and black people lived. We are also reminded of the fact that in the early phase of the COVID-pandemic, UK ministers ordered to discharge thousands of patients from hospital and move them to care homes without testing them for COVID in order to vacate hospital beds and protect the National Health System from being overburdened. A few weeks later, almost 1,000 care homes reported an outbreak of COVID (Keen 2023, 119). More generally, efforts of closing borders in Europe and elsewhere to seal oneself off from people seeking refuge from the disasters in their home country means that 'threatened populations are turned into threatening populations' (Keen 2023,46). Henry Giroux speaks of a 'new biopolitics of disposability': vulnerable populations like the poor, people of colour, the elderly, or people with disabilities, 'not only have to fend for themselves in the face of life's tragedies but are also supposed to do it without being seen by the dominant society.' (Giroux 2006, 175)

Coming back to digital disaster twins, I suggest we can derive a number of questions from these considerations: Who defines the disaster or emergency and how? In other words: Will decision-making about whether and when an intervention is necessary and what kind of intervention it should be, be automated? If so, who is accountable for the decision and its outcomes? Who determines whether in retrospect the intervention has been helpful and legitimate? Will there be any pre-determined priorities, rules or standards that would define the scope of possible decisions and interventions, or should decision-making be left entirely to the discretion of the -cyber-physical system in order to

allow it to respond optimally to unforeseen and unprecedented situations? After all, it is impossible to foresee all possible situations of disaster in advance and define a pre-selected set of appropriate interventions. Therefore, it is tempting to hope that automated decision-making on the basis of as complete as possible information, based on real-time data, will allow more immediate, flexible, adequate and efficient decisions than any human operator could take.

If code is law (Lessig 1999), we could say the digital disaster twin is sovereign. Sovereign, to paraphrase Carl Schmitt (Schmitt 2005), is the one who decides on the state of emergency, meaning that the rules and norms governing in normal times are suspended and measures and interventions that would normally be against the law are allowed in order to fight off an extraordinary situation that threatens to undermine the existence of the state. The self-preservation of the system rules supreme and overrides the rule of law. After all, the rationale for the sovereign power to suspend the law is that no law-maker, however wise and knowledgeable, can ever anticipate all possible emergencies that could require state intervention. Therefore, so the case for the state of exception, any law would unduly constrain the state's capacity to defend itself in a situation of emergency and safeguard its very existence. By way of analogy, the digital disaster twin, along this rationale, needs to be unbound by predefined laws and rules in order to be able to effectively fight off any possible existential threats to the system it is supposed to preserve. The problem is, as we know, that the state of emergency tends to become the norm itself and the sovereign tends to cling on to the powers of emergency rather than returning to the rule of law (Agamben 2005, Rossiter 1963[1948]).

If we take the dynamic digital twin to be a cyber-physical system capable of self-management, self-regulation and self-preservation, the question therefore is: What is the system that is to regulate, stabilise and preserve itself? Does self-preservation include stabilising the fault lines of inequality, powerlessness, marginalisation and exclusion that render certain groups vulnerable in the first place, for instance by trapping them in areas affected by floods, draught, viruses or other disasters in order to protect the other parts of the city from their presence? If the city is the system to be preserved and the city is already structured by fault lines of 'disposability', as Giroux puts it, preserving the city by preserving disposability is not a contradiction in terms. Systemic self-regulation by a digital twin may well mean to protect some lives at the expense of others.

To conclude, digital disaster twins could become self-regulating cyber-physical systems that at best reiterate the existing problems of disaster management as outlined above, depoliticising the definition of disaster as well as the decisions about the interventions taken and concealing the social and political causes of disasters in the first place. In the worst case, a digital disaster twin would automate the biopolitics of disposability.

What would it take to prevent digital disaster twins from exacerbating the problems of disaster and emergency management outlined above? What would it take to prevent

digital disaster twins from further depoliticising disaster management and its exclusionary, stratifying, biopolitical implications? What would it take to prevent digital disaster twins from automating and depoliticising a biopolitics of disposability? What kind of political debate, among whom, in which kind of arena and at which point in time could be conceived to deal with these issues?

References

Agamben, Giorgio. 2005. State of Exception. Chicago: The University of Chicago Press.

ARC. 2024. Atlanta Regional Commission. Digital Twins Whitepaper 2024. https://cdn.atlantaregional.org/wp-content/uploads/digital-twins.pdf.

Ariyachandra, M. R. Mahendrini Fernando and Gayan Wedawatta. 2023. 'Digital Twin Smart Cities for Disaster Risk Management: A Review of Evolving Concepts '. Sustainability 15(15):11910.

Casadei, Roberto, Danilo Pianini, Mirko Viroli and Danny Weyns. 2022. 'Digital Twins, Virtual Devices, and Augmentations for Self-Organising Cyber-Physical Collectives.' Applied Sciences 12(1):349. doi: https://doi.org/10.3390/app12010349.

Charitonidou, Marianna. 2022. 'Urban Scale Digital Twins in Data-Driven Society: Challenging Digital Universalism in Urban Planning Decision-Making.' International Journal of Architectural Computing 20(2):238-53.

Cheng, Ruijie, Lei Hou and Xu CHeng. 2023. 'A Review of Digital Twin Applications in Civil and Infrastructure Emergency Management.' Buildings 13(5).

Deren, Li, Yu Wenbo and Shao Zhengfeng. 2021. 'Smart City Based on Digital Twins.' Computational Urban Science 1:1-11.

Doğan, Ö., O. Şahin and E. Karaarslan. 2021. 'Digital Twin Based Disaster Management System Proposal: Dt-Dms.' Journal of Emerging Computer Technologies 1(2):25-30.

El-Agamy, Rasha F., Hanaa A. Sayed, Arwa M. AL Akhatatneh, Mansourah Aljohani and Mostafa Elhosseini 2024. 'Comprehensive Analysis of Digital Twins in Smart Cities: A 4200-Paper Bibliometric Study.' Artif Intell Rev 57(154).

Fan, Chao, Cheng Zhang, Alex Yahja and Ali Mostafavi. 2021. 'Disaster City Digital Twin: A Vision for Integrating Artificial and Human Intelligence for Disaster Management.' International Journal of Information Management 56(February 2021):102049.

Ferré-Bigorra, Jaume, Miquel Casals and Marta Gangolells. 2022. 'The Adoption of Urban Digital Twins.' Cities 131:103905.

Ford, David N. and Charles M. Wolf. 2020. 'Smart Cities with Digital Twin Systems for Disaster Management.' Journal of management in engineering 36(4).

Giroux, Henry A. 2006. 'Reading Hurricane Katrina: Race, Class, and the Biopolitics of Disposability.' College Literature 33(3):171-96.

Grieves, Michael. 2002. ‘Completing the Cycle: Using Plm Information in the Sales and Service Functions ‘ Paper presented at the SME Management Forum, South Orange, NJ.

Grieves, Michael. 2014. Digital Twin: Manufacturing Excellence through Virtual Factory Replication. A Whitepaper by Dr. Michael Grieves. https://www.researchgate.net/publication/275211047.

Grieves, Michael and John Vickers. 2017. ‘Digital Twin: Mitigating Unpredictable, Undesirable Emergent Behavior in Complex Systems.’ Pp. 85-113 in Transdisciplinary Perspectives on Complex Systems: New Findings and Approaches, edited by F.-J. Kahlen, S. Flumerfelt and A. Alves. New York: Springer.

Grieves, Michael. 2024. ‘Intelligent Digital Twins and the Development and Management of Complex Systems.’ Digital Twin 1(1):1-24. doi: https://doi.org/10.12688/digitaltwin.17574.1.

Keen, David. 2023. When Disasters Come Home. Cambridge, UK: Polity.

Kephart, Jeffrey O. and David M. Chess. 2003. ‘The Vision of Autonomic Computing.’ Computer 36(1):41-50. doi: 10.1109/MC.2003.1160055.

Kitchin, Rob and Oliver Dawkins. 2024. ‘Digital Twins and Deep Maps.’ Transactions 50(1):e12699.

Korenhof, Paulan, Vincent Blok and Sanneke Kloppenburg. 2021. ‘Steering Representations—Towards a Critical Understanding of Digital Twins.’ Philosophy & Technology 34:1751-73.

Korenhof, Paulan, Else Giesbers and Janita Sanderse. 2023. ‘Contextualizing Realism: An Analysis of Acts of Seeing and Recording in Digital Twin Datafication.’ Big Data & Society 10(1):1-12.

Lagap, Umut and Saman Ghaffarian. 2024. ‘Digital Post-Disaster Risk Management Twinning: A Review and Improved Conceptual Framework.’ Reduction 110:104629.

Lessig, Lawrence. 1999. Code and Other Laws of Cyberspace. New York, NY: Basic Books.

Liu, Chuncheng and Ying Tian. 2023. ‘Recognition of Digital Twin City from the Perspective of Complex System Theory: Lessons from Chinese Practice.’ Journal of Urban Management 12(2):182-92.

Martinescu, Livia. 2023. ‘Exploring the Concepts of Digital Twin, Digital Shadow, and Digital Model.’ Oxford Insights (23 October 2023).

Mohiuddin Eumi, Ettilla. 2024. ‘A Systematic Review of Digital Twins in Efficient Pandemic Management with Challenges and Emerging Trends.’ Decision Analytics Journal 12:100502.

Mylonas, Georgios, Athanasios Kalogeras, Georgios Kalogeras, Christos Anagnostopoulos, Christos Alexakos and Luis Muñoz. 2021. ‘Digital Twins from Smart Manufacturing to Smart Cities: A Survey.’ IEEE Access 9:143222-49.

Pickering, Andrew. 2010. The Cybernetic Brain. Chicago: University of Chicago Press.

Rocha, Andre Dionisio and Jose Barata. 2021. ‘Digital Twin-Based Optimiser for Self-Organised Collaborative Cyber-Physical Production Systems.’ Manufacturing Letters 29(August):79-83.

Rossiter, Clinton. 1963[1948]. Constitutional Dictatorship. Crisis Government in the Modern Democracies. New York: Harcourt, Brace & World.

Rothe, Delf. 2024. ‘When the World Is an Object: On the Governmental Promise of a Digital Twin Earth.’ International Political Sociology 18(olae022).

Schmitt, Carl. 2005. Political Theology. Four Chapters on the Concept of Sovereignty. Translated and with an Introduction by George Schwab. With a New Foreword by Tracy B. Strong. Chicago: The University of Chicago Press.

Sepasgozar, Samad M. E. 2021. ‘Differentiating Digital Twin from Digital Shadow: Elucidating a Paradigm Shift to Expedite a Smart, Sustainable Built Environment ‘. Buildings 11(4).

Tomko, Martin and Stephan Winter. 2019. ‘Beyond Digital Twins – a Commentary.’ Environment and Planning B: Urban Analytics and City Science 46(2):395–99. doi: https://doi.org/10.1177/2399808318816992.

WEF. 2022. Digital Twin Cities: Framework and Global Practices. Insight Report: World Economic Forum.

WEF. 2023a. World Economic Forum in Collaboration with the China Academy of Information and Communications Technology (Caict): Digital Twin Cities: Key Insights and Recommendations. Insight Report August 2023. Geneva: World Economic Forum.

WEF. 2023b. How Digital Twins - and Metavercities - Could Reshape Urban Development from Auckland to Hong Kong. ttps://www.weforum.org/stories/2023/06/digital-twin-city-metavercity-auckland-hong-kong/.

Willige, Andrea. 2022. Digital Twins: What Are They and Why Do They Matter? May 24, 2022. https://www.weforum.org/stories/2022/05/digital-twin-technology-virtual-model-tech-for-good/.

Zio, Enrico and Leonardo Miqueles. 2024. ‘Digital Twins in Safety Analysis, Risk Assessment and Emergency Management.’ Reliability Engineering & System Safety 246(110040):1-13.

The Future of Digital Humanism – Towards a Critical Post-Post-Humanism

Katja Mayer[1], Erich Prem[2], Philip Birkner[3], Pia-Zoe Hahne[4], Alexander Schmölz[4], Francesco Striano[5], Maria Zanzotto[5]

[1] University of Vienna and Center for Social Innovation ZSI Austria

[2] eutema GmbH and Digital Humanism Association, Vienna, Austria

[3] Independent Expert in Digital Transformation Policy, Vienna, Austria

[4] UAS BFI Vienna, University for Economics, Management and Finance, Austria

[5] University of Turin, Italy

DOI 10.3217/978-3-99161-062-5-002, CC BY 4.0

Abstract. This position paper distills key insights from the STS Graz 2025 panel 'The Future of Digital Humanism: Towards a Critical Post-Post-Humanism?'. The session brought together interdisciplinary perspectives to discuss Digital Humanism in light of feminist, ecological, infrastructural, and socio-economic critiques. While the movement draws on humanist ideals like dignity and autonomy, panelists emphasized the need to move beyond Western-centric and techno-solutionist narratives. They proposed a pluralistic and situated approach, framing Digital Humanism as a boundary object - flexible across contexts yet grounded in shared normative orientations. The paper outlines five theses: (1) Digital Humanism should not be equated with classical humanism but understood as a political response to digital dehumanization; (2) critical engagement with humanism helps to resist the powerful narratives of determinism and integrating situated epistemologies and feminist STS helps avoid universalist assumptions and centers marginalized perspectives; (3) more inclusive and accountable digital futures require sustained political engagement and the development of public digital infrastructures; (4) trust in generative AI needs to be reframed as a critical and reflective practice; (5) ecological responsibility can be strengthened through relational ethics that tie human well-being to environmental sustainability. In conclusion, translating theory into practice calls for institutional support and collaborative communities of action across disciplines and sectors. Together, these contributions reimagine Digital Humanism as an evolving, practice-oriented framework - capable of engaging diverse knowledge traditions while promoting democratic, just, and ecologically sound responses to digital transformation.

1 Introduction

This position paper presents core insights from the panel session with the provocative title ‘The Future of Digital Humanism: Towards a Critical Post-Post-Humanism?’, held at STS Graz Conference in May 2025.

This session convened a range of critical perspectives that examined the evolving contours of Digital Humanism through feminist, ecological, economic, educational, and infrastructural lenses. The panel’s provocation - whether Digital Humanism requires a ‘post-post-humanist’ turn - was addressed through a productive tension: rather than abandoning humanism as outdated, contributors explored how its core values might be re-situated, expanded, and politicized to meet contemporary socio-technical challenges.

Despite differences in philosophical orientation, the contributors shared a deep concern for the prevailing limitations of both techno-utopianism and Western-centric narratives that often masquerade as universal human-centrism. Instead, they called for a re-articulation of human-technology relations - one that is attentive to plural epistemologies, situated knowledge, and the material and institutional conditions shaping digital life-worlds. The discussion illuminated the generative momentum of the Digital Humanism movement and emphasized the imperative to transform this momentum into tangible inter- and transdisciplinary collaboration. Such an approach is essential not only for anchoring shared normative commitments across diverse fields, but also for shaping the ethical, political, and infrastructural frameworks through which Digital Humanism might intervene in the digital condition (Stalder, 2018) of our time.

Digital Humanism as defined in the ‘Vienna Manifesto on Digital Humanism’ is a movement and philosophical approach that emphasizes placing human values, rights, and dignity at the center of technological development, particularly in the age of digital transformation and Artificial Intelligence (Werthner, 2025). It calls for critically examining how digital technologies shape society and aims to ensure that these tools serve democratic principles, individual freedoms, and social cohesion rather than undermining them.

According to its advocates, Digital Humanism must actively guide technological innovation to align with societal goals. In this view, the history of Digital Humanism is rooted in European intellectual traditions that champion human agency, reason, and ethics, now confronted with the accelerating impact of algorithms and automation. Digital humanism has argued that digital technologies must be transparent, aligned with societal values including sustainability, diversity, and consideration of non-humans. It needs to be developed with interdisciplinary insight, including from the social sciences and humanities. Digital Humanism is not just a critique of Big Tech’s power but a constructive call to redesign systems where technology empowers rather than controls people. Importantly, this call was very much driven by computing professionals, hence it emerged

not as a call to theoretical deliberation, but as a move to a changing practice of IT design and education.

The five theses presented here emerged through collaborative post-panel synthesis. Following the session, contributors identified points of convergence across their presentations, focusing on recurring themes, productive tensions, and shared normative commitments. Selection criteria prioritised arguments that addressed limitations in existing Digital Humanism discourse, drew on distinct disciplinary perspectives whilst remaining in dialogue with one another, and offered both critical and constructive orientations. The resulting theses represent neither a simple aggregation of individual views nor a consensus document, but a distillation of what the contributors jointly recognised as essential coordinates for a critically grounded Digital Humanism. Our contribution is thus integrative rather than paradigm-shifting: we seek to enrich and re-situate Digital Humanism through critical perspectives often underrepresented in its foundational texts, not to replace it with an altogether different framework.

2 Five Theses for Digital Humanism

Thesis 1: Digital Humanism is not to be reduced to humanism.

While Digital Humanism draws from Renaissance and Enlightenment ideals - such as human dignity, autonomy, and rationality - it cannot be equated with classical humanism, nor should it be caricatured as merely a Western, anthropocentric project. Rather, Digital Humanism should be understood as a political and ethical response to contemporary forms of digital anti-humanism: the socio-technical processes through which algorithmic systems, data-driven infrastructures, and platform economies erode human agency, reduce persons to data points, and perpetuate systemic exclusion and control.

Digital Humanism, in this light, is a pragmatic and pluralist framework for resisting such dehumanization. It insists on re-centering the human - not as a fixed essence, but as a political figure of concern - to confront the loss of meaningful participation, accountability, and justice in the digital condition. This project calls for philosophical responsibility: acknowledging the critiques of humanism from feminist, postcolonial, and posthumanist perspectives without abandoning its historical democratic and emancipatory ambitions. These critiques have exposed how historical humanism often centered a narrow, Eurocentric ideal of the human - one that excluded women, racialized people, non-Western worldviews, and non-human life. Yet rejecting humanism wholesale risks discarding its potential as a framework for solidarity and normative orientation.

What is needed is a dialectical re-engagement with humanism: one that does not universalize from the particular, but instead understands the universality - e.g. of our fundamental rights - as an ongoing negotiation across differences. Humanity, in this view,

is not a fixed endpoint but an open, evolving condition shaped through the interplay of diverse perspectives, histories, and experiences. What we call ‘the human’ emerges in the space of communication and mutual recognition - not as a static identity, but as continually formed and reformed through relational encounters. In this sense, humanity is not a preordained endpoint, but a shared, unfolding condition (Collège des Bernardins, 2024), or a common ground (Arendt, 2010).

This has profound implications for the digital environment. The interconnectedness and ubiquity of digital infrastructures produce new forms of technological universalism - whether through globally shared platforms, interoperable standards, or more generally the planetary reach of computation and its power concentrations. While some digital phenomena can be localized, others exert effects that are necessarily transnational and transcultural and must be addressed through context-sensitive common grounds. Digital Humanism, then, requires tools to distinguish between harmful universalism and a pluralist, communicative universality grounded in situated yet interrelated experiences. Rather than opposing universality outright, the task is to shape it ethically: not as abstraction, but as a practice of relational inclusion - capable of guiding political agency, institutional design, and collective responsibility in the digital age (Prem, 2024). As a term, digital humanism is misleading as it never originated from an essentialist conception. It aims at a praxis and societal movement that recognizes humans and societies in their diversity, fosters regeneration and environmental views.

Thesis 2: Digital Humanism resists determinism and requires situated epistemologies.

Digital Humanism offers a critical alternative to two dominant yet problematic narratives of digital transformation: techno-determinism and techno-solutionism. Against the former, it challenges the belief that technological development unfolds according to its own inevitable logic - driven solely by efficiency, scalability, or market rationality - rendering societies passive and reactive (Winner, 1985; Wyatt, 2008). Such a view erases human agency, forecloses democratic deliberation, and normalizes the idea that ‘there is no alternative’ to technological trajectories set by corporate or state actors. In contrast, Digital Humanism re-centers society as a political and epistemic agent in the co-creation of technology. It emphasizes that digital futures are not predetermined but open to contestation, redirection, and collective shaping. This orientation is essential for those seeking not only to critique existing systems but to engage in transformative practice.

At the same time, Digital Humanism resists techno-solutionism - the belief that complex social problems can be solved primarily through technological innovation. This perspective treats technologies as neutral tools and obscures the political, economic, and cultural dimensions of both the problems and their supposed solutions (Eubanks, 2018; Morozov, 2013). Digital Humanism rejects the notion that technological artifacts are neutral; they embody and reinforce specific values, biases, and institutional agendas.

Rather than placing blind faith in innovation, it advocates a politics of bounded optimism - supporting technological advancement while foregrounding the need for deliberation, contextual sensitivity, and political imagination. By resisting both the fatalism of technological determinism and the oversimplifications of techno-solutionism, Digital Humanism thus seeks to build upon and support those already engaged in practices such as participatory design, ethical reflection, and inclusive governance, working toward technologies that serve democratic and socially just aims.

A key strategy in this effort is to foster transdisciplinary collaboration - bringing together expertise from the humanities, social sciences, technical disciplines, and civil society (Mayer and Strassnig, 2020; Werthner et al., 2022). Countering determinism requires more than critique: it demands spaces where engineers, designers, policymakers, and affected communities can co-develop frameworks that situate technical decisions within their broader societal contexts. Such collaboration makes it possible to reframe questions not only around what can be built, but what should be built, for whom, and under what conditions. Digital Humanism, in aligning with these efforts, strengthens its role as a mediating practice, connecting critical reflection with real-world interventions.

For Digital Humanism to realize its democratic and inclusive ambitions in practice, it must critically examine its own epistemic foundations by engaging more deeply with insights from feminist epistemology and Science and Technology Studies (STS). Without this reflection, Digital Humanism would risk reproducing Enlightenment-derived ideals, such as autonomy, rationality, and dignity, as if they were ahistorical and apolitical. These values, although normatively important, have also functioned as mechanisms of exclusion, legitimizing the marginalization of those deemed ‘less human’ or ‘less rational’ within sociotechnical systems. Feminist and STS perspectives offer tools to interrogate such assumptions through concepts like situated knowledge, relational autonomy, and systemic responsibility. They challenge the binary of ‘humans vs. technology’ and expose the embedded power dynamics within digital infrastructures (Benjamin, 2019; Haraway, 1988; Harding, 1991). The concept of knowledge as situated further challenges the pretence of a disembodied, universal ‘view from nowhere,’ arguing that all knowledge is produced from particular locations and that acknowledging this partiality enables more accountable and responsible knowledge claims (Haraway 1988).

Situated epistemologies thus deepen this vision by grounding universality in the material, historical, and relational conditions of knowledge production. They emphasize that what is commonly called ‘the human’ is not a static essence, but an evolving horizon shaped through interdependent experiences and positionalities. To move beyond ethical abstraction, Digital Humanism must also adopt a context-sensitive, practice-based ethics - one that asks: Whose values are encoded in digital systems? Who defines dignity, autonomy, or justice in specific technological contexts? Which systemic inequalities are embedded? (Barocas and Boyd, 2017; Birhane et al., 2022; Crawford, 2021). Engaging with critical technoscience provides conceptual and operational tools, such as

participatory design, algorithmic accountability, and plural epistemologies that can transform normative commitments into grounded interventions (D'Ignazio and Klein, 2019; Green, 2021; Schäfer et al., 2024). This re-situation allows Digital Humanism to shift from declarative ideals to institutional imagination, where values are tested, negotiated, and enacted in real-world settings.

Such an orientation also broadens the capacity for inter- and transdisciplinary collaboration by creating shared vocabularies and practices across fields. It fosters participatory engagements with diverse stakeholders - including technical experts, social scientists, activists, policy makers and affected communities - to co-create more just and contextually attuned technologies. In doing so, Digital Humanism shifts from a normative stance to a situated mode of sociotechnical world-making - a generative practice grounded in democratic engagement, epistemic plurality, and collaborative reconfigurations of technology and society.

Thesis 3: Digital Humanism contributes to the development of robust politics.

Digital Humanism aims at overcoming the false binary of innovation versus regulation. It aims to foster a more dynamic dialogue between regulators, public institutions, and industry to reimagine the governance of digital technologies. One important pillar for this is regulation, which is currently under attack, as highlighted in the Digital Humanism's 'Open Letter on the Urgent Need to Regulate Digital Technologies' (Digital Humanism, 2024). However, the challenges we face cannot be met by regulatory instruments alone. While frameworks like the European General Data Protection Regulation GDPR provide necessary baselines, they are insufficient to address the structural inequities and extractive dynamics embedded in the digital economy. Regulation should be regarded as a foundation, not a ceiling, for more ambitious political and institutional transformations. This requires moving from defensive measures to constructive alternatives: not only regulating against harm, but building public infrastructures, inventing new institutions, and articulating digital rights as positive liberties.

To build robust digital politics, Digital Humanism must advocate for public digital infrastructures and commons-based alternatives that resist the dominance of market-driven logics. Rather than accepting the primacy of hyperscalers and extractive platform capitalism, we need coordinated public investment in ethical, accountable, and socially beneficial technological systems. Diverse public voices, such as Mariana Mazzucato and Evgeny Morozov, alongside the Draghi Report's call for strategic state intervention, point to the urgency of reasserting public agency in digital development (Draghi, 2024; Mazzucato et al., 2022; Morozov, 2013). Federated infrastructures, such as those envisioned by initiatives like the Open Future Foundation or the Next Generation Internet (NGI) Digital Commons, offer concrete models for how this could be realized, alongside already existing initiatives such as the Barcelona Decidim platform for participatory governance or public library consortia providing open-access digital services

(Barandiaran et al. 2024; Bosman et al. 2021; EC 2025; OF 2024). In the future, large language models (LLMs), for example, could be hosted as public infrastructures by public libraries or universities, serving society as knowledge commons and supporting linguistic minorities rather than as proprietary tools governed by opaque corporate interests (Samdub, 2025; Sieker et al., 2025).

This shift may require not just better tools, but also new institutions. This may require the invention of new and potentially digital participatory governance models and infrastructure designed around public interest values. These include community-based platforms, transparent AI oversight mechanisms, and state-led stewardship of data and computation as public goods. The environmental and social costs of extractive digital infrastructures, such as data centers and cloud computing, must be properly accounted for and no longer externalized. A fact-based understanding of these costs is essential to designing policies that can redistribute power and resources more equitably.

Fundamentally, Digital Humanism should promote a vision of digital positive liberties: not only the right to be protected from harm online, but the right to access, shape, and co-govern the digital tools and infrastructures that affect our lives. The call for ‘basic digital services’ to be recognized and treated as public infrastructure is gaining traction - and should be amplified by the Digital Humanism movement. The current trajectory of digital governance continues to marginalize civil society voices, even as the influence of large technology corporations expands. This imbalance poses a real threat to democratic legitimacy. Digital Humanism must thus serve as a normative and practical force to help correct it by advancing institutional innovations that center justice, participation, and the common good.

At the same time, Digital Humanism needs to also move beyond regulation. The dominance of parasitic, asset-intensive platform models and the entanglement with deregulated capital markets have created a form of techno-feudalism that regulatory competition policies alone cannot dismantle (Morozov, 2022). Besides digital services for the public good, we may need other concepts from a renewed entrepreneurial digital virtue to communities of ethical practices. Digital Humanism should therefore advocate for new institutions, federated digital public infrastructures and their governance models, and the better recognition of digital civil rights based on the fundamental rights to ensure equitable access and participation in the digital sphere.

Thesis 4: Trust in generative AI must be critically redefined.

The challenge of trust in generative AI exemplifies the epistemic and political stakes outlined in the preceding theses: here, the need for situated knowledge, critical literacy, and democratic accountability converges in a concrete domain of human-technology relations. Probabilistic systems like large language models challenge both traditional and emerging conceptions of epistemic trust. Drawing from a non-standard account by

Francesco Striano (Striano, 2024), trust is not simply a matter of belief or reliance but an evaluative act - a judgment about the trustworthiness of a system based on its perceived reliability. Under this model, it is conceivable to extend trust to technologies if they meet such evaluative criteria. However, with LLMs, trust is often granted without this reflective process. These systems produce outputs that appear coherent, authoritative, and reliable, even though they are generated through probabilistic mechanisms rather than deterministic reasoning. Their fluency and speed simulate reliability, fostering a misplaced trust that is more about user perception than about actual trustworthiness.

This misplaced trust is amplified by the interactive, human-like design of LLM-based chatbots, which encourages users to relate to them as if they were rational agents. Rather than cultivating critical scrutiny, the design of these systems often promotes uncritical engagement. Digital Humanism must counter this tendency by promoting epistemic agency and critical digital literacy. Users should be equipped to interrogate and contextualize AI-generated content, recognizing the narrative and probabilistic nature of these outputs rather than accepting them at face value. Critical digital literacy can be fostered at multiple stages of education and life - as exemplified by Finland's national media literacy curriculum (Salomaa & Palsa 2019) or emerging university programmes on algorithmic accountability - but it should not be relegated to individual responsibility, in contrast, design choices should promote a more critical interaction. For example, some researchers have experimented with different interfaces of a chatbot, primarily proving more than one answer to a single user's request; this revealed the stochastic nature of LLM and led users to engage critically with the chatbot and its outputs, counteracting blind trust (Swoopes, Holloway and Glassman, 2025). In this context, trust should not be understood as passive acceptance, but as a reflexive and evaluative stance that allows for dissent, doubt, and revision.

What we face today is a trust paradox: many users engage with AI systems not because they trust them in an epistemic or moral sense, but because of convenience, fear of missing out, or the belief that these technologies will improve over time. This habituated reliance is often mistaken for trust but lacks the core ingredients of moral evaluation and freely given consent. In the absence of viable alternatives or adequate transparency, users develop a form of pseudo-trust - confidence without critique - which ultimately erodes their epistemic autonomy and reinforces the illusion of AI's infallibility.

Traditional and feminist theories of trust - such as those advanced by Baier (1986) and Govier (1992) - emphasize trust as a relationship between moral agents, grounded in goodwill, mutual vulnerability, and normative commitment. Extending such conceptions of trust to AI systems risks conferring moral agency upon machines, thereby distorting the very idea of trust. Moreover, the harms introduced by AI are often systemic and collective, disproportionately affecting marginalized communities. As Smuha (2021) notes, these harms do not stem from individual betrayal but from the structural vulnerabilities perpetuated by the systems themselves. Digital Humanism, therefore,

must resist anthropomorphizing AI and instead build frameworks that support critical engagement, protect vulnerable populations, and uphold human freedom in the digital age.

Thesis 5: Critical and ecological notions substantiate existing understandings of Digital Humanism.

While Thesis 2 addressed the epistemological foundations of Digital Humanism, the question of our ethical relations to the non-human world requires a distinct treatment: not only how we know, but how we ought to relate. The intensifying discourse on digitalisation and its environmental footprint has led to a renewed interest in ethical frameworks guiding technological development. Digital Humanism (Doueihi, 2011; Nida-Rümelin and Weidenfeld, 2022; Werthner et al., 2022) offers a perspective by placing human beings - along with their capacity for ethical judgement, boundary-setting, co-creative engagement and multi-modal literacy (Schmoelz, 2020) - at the center of technology use and development. However, in the light of the critique it faces for its western- and male-centrism in the shape of human-centrism, we suggest that Digital Humanism can evolve into a more relational and ecologically aware ethics that responds to the ambivalent legacy of Enlightenment thought and aligns human well-being with responsible environmental stewardship. This does not require abandoning its human-centered focus but demands a rethinking of human-nature relations in relational rather than hierarchical terms. Relational ethics recognizes that human well-being is foundationally intertwined with the protection of the natural world. Relational ethics allow us to take into account different life worlds where not only human actors are recognised as active moral subjects, but instead also take non-human actors such as plants and nature as deserving of moral consideration (Coeckelbergh, 2018; Metz & Miller, 2013). Relational ethics allows us to inform our relationship with other entities, such as nature, meaning that in the end, it focuses on ‘how we humans, as relational beings, relate and are related to in general’ (Coeckelbergh, 2018, p. 106). We forward a critical and ecological Digital Humanism that critically emphasizes the rooting of exploiting nature and humans alike in the capitalist and progressive neoliberal project (Fuchs, 2022; Schmoelz, 2023).

The care for nature is rooted in the necessity of maintaining environmental conditions that enable human well-being as end-in-itself. Digital Humanism does not propose a purely eco-centric stance - an approach that might prioritize nature even at the expense of human well-being - but rather emphasizes a critical human-centered environmental ethic. As critics point out (Brevini, 2022; Lucivero, 2020; Saenko, 2023), the massive ecological costs of digital infrastructures, particularly AI technologies, pose urgent ethical questions. Digital Humanism addresses the vast energy demands of AI as well as the broader spectrum of environmental exploitation: the extraction of rare minerals, the consumption of water resources, and the global supply chains that support data-centers and device production. To advance a meaningful ecological engagement, Digital

Humanism critically reflects on its historical roots. Digital Humanism today consciously distances itself from this aspect of Enlightenment thinking, which partially overwrote the emancipatory notion of humanist philosophy (Horkheimer and Adorno, 1969).

Digital Humanism actively integrates principles of care ethics and ecological responsibility into its normative framework. Inspired by recent critiques of hierarchical and exclusive anthropocentrism (Braidotti, 2013; Coeckelbergh, 2024; Goodley et al., 2020; Prem, 2024), Digital Humanism can contribute an alternative vision to both anti-humanist hierarchical relation between humans and with nature as well as a post-humanist depersonalisation of agency. Digital Humanism calls for an ethics of responsibility: developing technologies and infrastructures that respect ecological limits while fostering human wellbeing. This perspective encourages rethinking digital technologies as embedded within social and ecological systems. It promotes human traits such as setting values and boundaries, multimodal literacy and co-creativity (Schmoelz, 2020) for fostering a more reflexive understanding of our impacts on the environment. The Vienna Manifesto on Digital Humanism has already highlighted the need for democratic control and humane values in digital development; the next step is to explicitly extend this to ecological considerations. This proposed critical and ecological Digital Humanism seeks a synthesis: upholding human dignity and agency while acknowledging the material conditions that sustain human well-being. We advocate for policies and design principles that minimize resource consumption, avoid unnecessary extraction, and promote circular economies within digital production chains.

Synthesis: Digital Humanism becomes effective when enacted as an embedded, collective practice.

To be clear, these theses do not exhaust the critical perspectives needed - questions of labour, global inequalities in AI development, and disability remain vital and underexplored - but represent the convergences that emerged from our particular interdisciplinary encounter. However, our theses trace a coherent arc: from reframing Digital Humanism as a political rather than essentialist project, through the epistemological and institutional conditions for its realisation - resisting determinism, centring situated knowledge, building public infrastructures, cultivating critical trust, and extending ethical concern to ecological relations. Together, they articulate Digital Humanism not as a fixed doctrine but as a reflexive framework shaped by feminist, ecological, and democratic commitments.

Taken together, all our theses argue that realizing the transformative potential of Digital Humanism requires its enactment as an embedded practice. It must evolve beyond a theoretical or philosophical orientation and take shape as a lived, situated practice - one that is actively embedded in academic, civic, business and political institutions. This requires the cultivation of long-term communities of practice in which scholars, technologists, policymakers, activists, and citizens work collaboratively to co-develop

ethical, inclusive, and context-sensitive responses to digital transformation. These communities cannot be engineered top-down or imported wholesale; rather, they must grow dialogically, rooted in existing initiatives, local conditions, and diverse knowledge traditions.

Such an approach foregrounds sustained collaboration and mutual learning over isolated expertise. Session discussants have stressed that the future of Digital Humanism hinges not only on interdisciplinarity, but on the co-creation of shared infrastructures, vocabularies, and political imaginaries. Institutionalizing Digital Humanism through these practices enables a shift from normative ideals to operational frameworks that support ethical reasoning, democratic engagement, and social justice in technological development. It also helps bridge the structural divide between critique and implementation, ensuring that ethical reflection is not confined to the margins of design but integrated into its everyday operations. Ultimately, it is through these embedded and participatory modes of engagement that Digital Humanism can move from aspiration to action. Rather than remaining an abstract call for better technology, it becomes a platform for collective world-making.

3 Conclusion: Digital Humanism must grow as a pluralistic movement

Our initial provocation of ‘post-post-humanism’ does not claim a new paradigm but signals a double movement: taking seriously the critiques of classical humanism whilst refusing to abandon some of its normative and political resources. What emerges is not a resolution of these tensions but a Digital Humanism that remains attentive to its own exclusions. So, in closing, we argue that Digital Humanism can only fulfill its ethical and political promise if it evolves as a pluralistic movement and a dynamic ‘boundary object’ (Star and Griesemer, 1989). Adaptable across diverse contexts while grounded in shared commitments, boundary objects enable coordination without requiring full consensus. However, to translate Digital Humanism’s current momentum into transformative practice, it needs to both broaden and deepen: broaden by engaging with diverse epistemic traditions, worldviews, and local contexts beyond Anglo-European paradigms; deepen by articulating normative principles that are ethically robust without reverting to essentialist or universalist abstractions. Rather than acting as a vessel for fixed normative meanings, Digital Humanism should operate as a flexible framework that enables meaningful collaboration across disciplinary, institutional, and geopolitical divides, while maintaining enough coherence to sustain a shared orientation toward democratic, just, and humane digital futures. This requires resonance without rigidity. As humanist values are taken up in different settings, they must be rearticulated in ways that reflect situated knowledge, lived experience, and plural visions of the good life. Remaining true to its humanistic aspirations entails centering global plurality and fostering genuinely transnational dialogues. Resisting epistemic hegemonies - especially those rooted in

dominant intellectual traditions - is not a rejection but an act of inclusion, necessary for co-creating ethical digital worlds in which multiple futures can be imagined and realized. The five theses presented in this paper aim to chart a path toward such a re-situated Digital Humanism - one that is critically grounded, politically engaged, and capable of shaping inclusive and sustainable digital futures.

References

Arendt, H., 2010. The human condition, 2. ed., [Nachdr.]. ed. Univ. of Chicago Press, Chicago.

Baier, A., 1986. Trust and antitrust. Ethics 96, 231-260.

Barandiaran, X.E., Calleja-López, A., Monterde, A., Aragón, P., Linares, J., Romero, C., Pereira, A., 2024. Decidim: Political and Technopolitical Networks for Participatory Democracy. Decidim Association, Barcelona. https://docs.decidim.org/en/develop/whitepaper/what-is-decidim.html Accessed 1.12.2025

Barocas, S., Boyd, D., 2017. Engaging the ethics of data science in practice. Commun. ACM 60, 23-25. https://doi.org/10.1145/3144172

Benjamin, R., 2019. Race after technology: abolitionist tools for the new Jim code. Polity, Medford, MA.

Birhane, A., Kalluri, P., Card, D., Agnew, W., Dotan, R., Bao, M., 2022. The Values Encoded in Machine Learning Research. https://doi.org/10.48550/arXiv.2106.15590

Bosman, J., Frantsvåg, J.E., Kramer, B., Langlais, P.-C., Proudman, V., 2021. The OA Diamond Journals Study: Exploring collaborative community-driven publishing models for Open Access. Part 1: Findings. Science Europe / cOAlition S. https://doi.org/10.5281/zenodo.4558704

Braidotti, R., 2013. The Posthuman. Polity Press, Cambridge.

Brevini, B., 2022. Is AI Good for the Planet? Polity Press.

Coeckelbergh, M. (2018). What do we mean by a relational ethics? In *Plant Ethics: Concepts and Applications* (1st edn, pp. 98-109). Routledge. https://doi.org/10.4324/9781315114392

Coeckelbergh, M., 2024. What is digital humanism? A conceptual analysis and an argument for a more critical and political digital (post)humanism. Journal of Responsible Technology 17, 100073. https://doi.org/10.1016/j.jrt.2023.100073

Collège des Bernardins, 2024. For a Critical Digital Humanism. Département Humanisme Numérique.

Crawford, K., 2021. The atlas of AI: Power, politics, and the planetary costs of artificial intelligence. Yale University Press.

Digital Humanism, 2024. An Open Letter on the Urgent Need to Regulate Digital Technologies [WWW Document]. DIGHUM. URL https://caiml.org/dighum/open-letter-on-the-urgent-need-to-regulate-digital-technologies/ (accessed 6.4.25).

D’Ignazio, C., Klein, L., 2019. Data feminism. MIT Press.

Doueihi, M., 2011. Pour un humanisme numérique, La librairie du XXIe siècle. Éditions du Seuil, Paris.

Draghi, M., 2024. The future of European competitiveness. EU Commission.

EC European Commission, 2025. Next Generation Internet Initiative. Shaping Europe's Digital Future. https://digital-strategy.ec.europa.eu/en/policies/next-generation-internet-initiative Accessed 1.12.2025

Eubanks, V., 2018. Automating inequality: How high-tech tools profile, police, and punish the poor. St. Martin’s Press.

Fuchs, C., 2022. Digital Humanism: A Philosophy for 21st Century Digital Society, SocietyNow. Emerald Publishing Limited, Bingley, UK.

Goodley, D., Lawthom, R., Liddiard, K., Runswick-Cole, K., 2020. The Desire for New Humanisms. Journal of Disability Studies in Education 1, 125-144. https://doi.org/10.1163/25888803-00101003

Govier, T., 1992. Trust, distrust, and feminist theory. Hypatia 7, 16-33.

Green, B., 2021. Data science as political action: Grounding data science in a politics of justice. Journal of Social Computing 2, 249-265.

Haraway, D., 1988. Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective. Feminist Studies 14, 575-599.

Harding, S., 1991. Whose science? Whose knowledge?: Thinking from women’s lives. Cornell University Press.

Horkheimer, M., Adorno, T.W., 1969. Dialektik der Aufklärung. Philosophische Fragmente. S. Fischer Verlag, Frankfurt a. Main.

Lucivero, F., 2020. Big Data, Big Waste? A Reflection on the Environmental Sustainability of Big Data Initiatives. Science and Engineering Ethics 26, 1009-1030. https://doi.org/10.1007/s11948-019-00171-7

Mayer, K., Strassnig, M., 2020. The Digital Humanism Initiative in Vienna: A Report based on our Exploratory Study Commissioned by the City of Vienna, in: Fritz, J., Tomaschek, N. (Eds.), Digitaler Humanismus. Waxmann Verlag. https://doi.org/10.5281/zenodo.4250144

Mazzucato, M., Schaake, M., Krier, S., Entsminger, J., 2022. Governing artificial intelligence in the public interest. UCL Institute for Innovation and Public Purpose, Working Paper Series (IIPP WP 2022-12). Retrieved April 2, 2023.

Metz, T., & Miller, S. C. (2013). Relational Ethics. In *International Encyclopedia of Ethics* (1st edn). Wiley. https://doi.org/10.1002/9781444367072

Morozov, E., 2022. Critique of Techno-Feudal Reason. New Left Rev 89-126.

Morozov, E., 2013. To save everything, click here: The folly of technological solutionism. PublicAffairs, New York.

Nida-Rümelin, J., Weidenfeld, N., 2022. Digital Humanism: For a Humane Transformation of Democracy, Economy and Culture in the Digital Age. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-031-12482-2

OF, Open Future Foundation, 2024. Digital Commons as Providers of Public Digital Infrastructures. Open Future, Amsterdam. https://openfuture.eu/publication/digital-commons-as-providers-of-public-digital-infrastructures/ Accessed 1.12.2025

Prem, E., 2024. Principles of digital humanism: A critical post-humanist view. Journal of Responsible Technology 17. https://doi.org/10.1016/j.jrt.2024.100075

Salomaa, S., Palsa, L., 2019. Media Literacy in Finland: National Media Education Policy. Ministry of Education and Culture, Helsinki. https://medialukutaitosuomessa.fi/mediaeducationpolicy.pdf Accessed 1.12.2025

Saenko, K., 2023. Is generative AI bad for the environment? A computer scientist explains the carbon footprint of ChatGPT and its cousins.

Samdub, M., 2025. Digital Public Infrastructure at a Turning Point.

Schäfer, M.T., Es, K., Lauriault, T., 2024. Collaborative Research in the Datafied Society. Collaborative Research in the Datafied Society.

Schmoelz, A., 2023. Digital Humanism, Progressive Neoliberalism and the European Digital Governance System for Vocational and Adult Education. Journal of Adult and Continuing Education. https://doi.org/10.1177/14779714231161449

Schmoelz, A., 2020. Die Conditio Humana Im Digitalen Zeitalter. Zur Grundlegung Des Digitalen Humanismus Und Des Wiener Manifests. MedienPädagogik. Zeitschrift Für Theorie Und Praxis Der Medienbildung 208-234. https://doi.org/10.21240/mpaed/00/2020.11.13.X

Sieker, F., Tarkowski, A., Gimpel, L., Osborne, C., 2025. Public AI - White Paper 73 S. https://doi.org/10.11586/2025040

Smuha, N.A., 2021. Beyond individual harm: Conceptualising AI's broader societal impacts, in: Dubber, M.D., Pasquale, F., Das, S. (Eds.), The Oxford Handbook of AI Ethics. Oxford University Press.

Stalder, F., 2018. The digital condition, English edition. ed. Polity Press, Cambridge Medford, [Massachusetts].

Star, S.L., Griesemer, J.R., 1989. Institutional Ecology, `Translations' and Boundary Objects: Amateurs and Professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39. Social Studies of Science 19, 387-420.

Striano, F., 2024. The Vice of Transparency. A Virtue Ethics Account of Trust in Technology.

Swoopes, C., Holloway, T., Glassman, E.L., 2025. The Impact of Revealing Large Language Model Stochasticity on Trust, Reliability, and Anthropomorphization, in arXiv.org, URL https://arxiv.org/abs/2503.16114 (accessed 10.10.25).

Werthner, H., 2025. Digital Humanism and the Vienna Manifesto, in: Digital Humanism. Springer Nature Switzerland, Cham, pp. 83-87. https://doi.org/10.1007/978-3-031-86905-1_7

Werthner, H., Prem, E., Lee, E.A., Ghezzi, C. (Eds.), 2022. Perspectives on Digital Humanism. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-030-86144-5

Winner, L., 1985. Do Artefacts Have Politics? In The Social Shaping of Technology: How the Refrigerator Got Its Hum, Donald MacKenzie and Judy Wacjman (eds.), 26-38.

Wyatt, S., 2008. Technological Determinism is Dead; Long Live Technological Determinism, in: Hackett, E., Amsterdamska, O., Lynch, M., Wajcman, J. (Eds.), The Handbook of Science and Technology Studies. MIT Press, Cambridge, MA, pp. 165-180.

Trust in Research Practices & Infrastructures

Katharina Flicker[1], Stefan Reichmann[2], Susanne Blumesberger[3], Marie Czuray[1], Miguel Rey Mazon[2], Bernd Saurugger[1], Andreas Rauber[1]

[1] TU Wien, Austria

[2] Graz University of Technology, Austria

[3] University of Vienna, Austria

DOI 10.3217/978-3-99161-062-5-003, CC BY 4.0

Abstract. The increasing complexity of digital research workflows raises questions about trust in research processes, results, and infrastructures. This study builds on philosophical concepts of trust to examine their relevance to research practices, particularly in relation to data, tools, services, and open-source software. We explore how trust influences sharing and reuse, the perception of quality indicators, and the development of trustworthy infrastructures. Two exploratory approaches were employed: a survey among data scientists in open-source software, and twelve semi-structured interviews with researchers from various disciplines focusing on trust in data quality. Interviews were transcribed and analysed using inductive coding supported by ATLAS.ti. Findings reveal a consistent gap between research ideals and practice. While researchers recognize the importance of verifying the fitness for purpose for reused resources, time constraints often lead them to rely on proxies such as documentation and source reputation. Trust is closely tied to institutional affiliation, peer review, and ethical standards, indicating that reputation and adherence to ethical codes influence perceptions of trustworthiness. The results also highlight the need for mechanisms to assess and communicate trustworthiness especially in dynamic and interdisciplinary contexts. Questions arise about integrating such mechanisms into research infrastructures, including standards for documentation, compliance monitoring, and responses to violations. This work lays the foundation for future research on institutional and technical processes that can foster trust and trustworthiness in the development and use of digital research infrastructures.

1 Introduction

Research is increasingly based on digital data, collected from a variety of sources, pre-processed and analysed by many stakeholders. The value of such and the need to protect the massive investments in terms of time, money and computing resources having gone into this, led to the building of complex data and compute infrastructures such as encountered in the proliferation of AI research and deployments. These are expected to ensure that research outputs are securely stored, properly documented, available for (re-)use, and compliant with the FAIR Principles[1] (Wilkinson *et al.*, 2016), which purport to streamline research data management practices to make digital research outputs (Wilkinson *et al.*, 2018) reusable by machines and humans (Boeckhout, Zielhuis, and Bredenoord, 2018).

Several scholars have proposed to study the complexity of data processing pipelines in terms of 'data journeys' (Leonelli, 2016; 2020) or 'data distance' (Borgman and Groth, 2025), to conceptually accommodate the amount of invisible labour and actors involved in rendering digital research outputs reusable (Leonelli, 2016). While this discussion is empirically saturated (cf. Leonelli and Tempini, 2020), the quality requirements for digital research outputs for a particular type of research, raise important (yet unanswered) questions concerning the trustworthiness of the underlying processes. More recently, the difficulties in assessing the quality of digital research outputs – ranging from specific findings to complex AI models – have spawned concerted efforts to tease apart the details of (input and result) data quality[2]. It is worth noting that FAIR constitutes a paradigm for guaranteeing the machine-readability of digital research outputs, while saying nothing about quality.

This article discusses the problems of ensuring quality of research outputs in terms of the trustworthiness of underlying processes, starting by classic and recent work on trust and trustworthiness in digital environments in sections 2.1, 2.2 and 2.3. It addresses recent academic crises (e.g., reproducibility) and their effects on trust in research processes, followed by an assessment of the infrastructural turn in STS in sections 2.3 and 2.4. While the research design is described in section 3, section 4 calls attention to empirical qualitative and quantitative data illustrating how researchers evaluate the trustworthiness of digital research outputs. Concluding remarks in section 5 highlight

[1] The FAIR principles mandate that research data be Findable, Accessible, Interoperable, and Reusable (Wilkinson et al., 2018), to ensure that digital research objects can be discovered and reused.

[2] Both RDA and the EOSC Association have invoked Task Forces (TF) and Working (WG) as well as Interest Groups (IG) to tackle these and related issues. The EOSC Association, for example, established the FAIR Metrics and Data Quality Task Force (2021-2023), and the FAIR Metrics and Digital Objects TF (2024-2025) (https://eosc.eu/eosc-association/eosc-task-forces/), while RDA runs the FAIR Mapping WG (https://www.rd-alliance.org/groups/fair-mappings-wg/activity/), or the FAIR Instrument Data IG (https://www.rd-alliance.org/groups/fair-instrument-data-ig/activity/) among many others.

implications for designing infrastructures and guidelines to enhance trust via mechanisms, processes, and human factors.

2 Related Research: Trust, credibility, and (research) infrastructures

This section begins by examining the philosophical foundations of trust and develops a working definition of the concept (2.1). It then illustrates how trust has been conceptualized and studied within the field of information and communication technology (2.2) and explores its role in research processes, outcomes, and infrastructures (2.3). Finally, drawing on insights from Science and Technology Studies, the section defines infrastructures and highlights the critical importance of addressing trust-related issues within the context of research infrastructures (2.4).

2.1 Trust

Trust is elusive. While the concept has been discussed in philosophy and related fields, there are many competing, often contradicting, definitions (McLeod, 2021). In social situations, there tends to be what Walker (2006) dubbed 'default trust' – based on a tacit understanding of what can be expected from one's surroundings. Barber (1983) provides a systematization of different aspects of trust, distinguishing 1) trust in the continuity of natural and social orders, 2) trust in the technical competence of an actor, and 3) trust in an actor's propensity to consider the interests of others (morality). Trust, then, is important across social settings. Social theorists recognize trust as a prerequisite of specifically modern (as opposed to traditional) forms of social interaction, by linking it to the contingency of social interaction: Hardin (2006; 2001), Gambetta (1988), and Luhmann (1979) observe that trust involves a moment of uncertainty about the behaviour of others. Contingency implies that actions can have unintended consequences (Merton, 1936); trusting others solves problems of contingency at the interpersonal, group, and societal levels (Luhmann, 1979). Similarly, Georg Simmel observed that trust is located, epistemically, between complete knowledge and complete ignorance of the other, and is therefore a background assumption of everyday life. In this, trust is essential for social interaction (Lewis and Weigert, 1985), and seems to be irreducibly social (Lewis and Weigert, 2012). Giddens (1996) later observed that trust serves to reduce the complexity of social interactions in that it reduces the 'costs' associated with having to verify others' intentions (Luhmann, 1979). Modern societies inherently rely on trust owing to increasing rationalization, which necessitates reliance on the expertise of others (Collins, 2007). For Luhmann (1979), trust sustains (relatively) stable expectations regarding the natural and social orders.

In philosophy, trust is the subject of theoretical and practical philosophy, where the concern is with finding an evidence-based and rational definition. These discussions are

too complex to track here, so the following working definition is proposed: Trust is warranted when it is plausible, well-grounded, and justified. Trust is plausible only if the conditions for trust are obtained, and when one is able to develop trust. Trust is well-grounded only if the trustee is trustworthy. It can be justified even if the trustee is not trustworthy as long as some value can be expected to emerge from giving trust (McLeod, 2021).

Defining trustworthiness is even more elusive, although there is a tendency to define the concept relative to trust. Hawley (2014) offers a helpful working definition based on three assumptions: 1) Trustworthiness is a trait or attribute of someone/something, 2), there is a difference between general and specific trustworthiness, and 3) to be trustworthy means to meet reasonable and appropriate expectations.

2.2 Trust in Information and Communication Technology (ICT)

There is a well-developed body of literature on trust in organisations, data, and technologies. With the growing amount of data, questions about how trust enhances its value arise in scientific and mundane contexts (Pink, Lanzeni and Horst, 2018). The relationship between trust in technology and its subsequent use is gaining significant attention. For example, Tronnier, Harborth and Hamm (2022) explore how privacy concerns and trust in currency affect individuals' willingness to adopt Central Bank Digital Currency; (Jacovi *et al*., 2021) study trust as a crucial component in human-AI interactions and in ICT more broadly (McKnight *et al.*, 2011).

With the wide-spread use of ICT, the notion of e-trust has been introduced to describe forms of ‘trust specifically developed in digital contexts and/or involving artificial agents’ (Taddeo and Floridi, 2011). It typically pertains to trust in online environments but has also been used to describe more general issues of digitally mediated social interactions. E-trust is typically analysed along the following three dimensions: 1) Trust in technologies, 2) trust in other users, and 3) trust in technology providers (Taddeo and Floridi, 2011), to which Spiekermann (2016) adds trust in the engineers who built a particular technology. Accounts of e-trust tend to have a cognitive leaning, frequently modelling e-trust as a relationship between a trustor and a trustee, because trust in artefacts cannot be based on either morality (Nickel, Franssen, and Kroes, 2010) or motivations (Hardin, 2006). As Sztompka (1999) points out, there might be a continuum from trust based on rationality to trust based on morality (i.e. rationality begets morality): Betting on virtue is riskier than betting on rationality. Taddeo (2010) suggests explicating trustworthiness not as general reliability, but rather as reliability with respect to performing a specific task. Trust is only achieved when coming from the right reasons: care for the other’s interest combined with moral integrity (Nickel, Franssen, and Kroes, 2010). These may be difficult to assess in digital environments. For this reason, Spiekermann (2016) proposes to speak of reliance on technology rather than trust, and correspondingly, reliability instead of trustworthiness.

2.3. Trust in, and credibility of, the research process

The following explores trust in research processes, results, and infrastructures, while not asking why the general public can and should trust science (Oreskes, 2019), but rather, which mechanisms (if any) warrant researchers to trust their own methods, tools, findings, etc. In this context, trust does not need to be absolute but must be justified and supported by evidence.

This understanding of trust is the kind that should be earned by research practices, infrastructures, and outcomes. An important precursor to this discussion is Merton's (Merton, 1973) work on 'The normative structure of science', which claims that in order for science to convey certified knowledge, researchers adhere to a set of four ('Mertonian') norms (Hosseini et al., 2024): universalism, communalism, disinterestedness, and organized scepticism, each with more or less bearing on the emergence of trust. In Merton's view, science is 'disinterested' in the sense that there are no 'external' customers, i.e., the consumers of scientific communication are other scientists. Organised scepticism entails that researchers may trust the system of peer review as a whole without having to check, for each research output, whether it is trustworthy – provided that they trust the system and understand that it has been peer reviewed.

This picture has come under heavy scrutiny (and criticism), under the impression of recent crises of academia, such as the crisis of peer review (Horbach and Halffman, 2018; Daniel, Mittag, and Bornmann, 2007; Smith, 2006), and – connected to the first – the reproducibility crisis (Ioannidis, 2005; Begley and Ioannidis, 2015; Leonelli, 2018; França and Monserrat, 2019), both of which tend to negatively impact the (perceived) trustworthiness of science.

Oreskes (2019) finds that trust in science is based upon consensus. Since there is no single valid scientific method, trust in its outcomes is based on the social character of the scientific process. The trustworthiness of science stems precisely from the social processes by which published results are discussed, reviewed, contested, and ultimately accepted in a 'cycle of credibility' (Latour and Woolgar, 1986).

It should be noted, that the process of allocating credibility in academia is subject to the 'Matthew effect' (Merton, 1968), a dynamic of cumulative advantage (Ross-Hellauer *et al.,* 2022) whereby early success (measured in terms of scholarly impact, i.e. citations) begets later success. Already successful researchers tend to receive rewards and recognition disproportionately, which tends to translate into resources and access to infrastructure, resulting in an extremely stratified distribution of these resources. For Merton, the Matthew effect fulfils an important function at the system level, in that it serves to assess the credibility of sources (similar to trust more generally). At the individual level, the effects of cumulative advantage – 'a general mechanism for inequality across any temporal process (e.g., life course, family generations) in which a

favourable relative position becomes a resource that produces further relative gains' (DiPrete and Eirich, 2006) – tend to be detrimental, as 'various aspects of academia are particularly vulnerable to logics of cumulative advantage' (Ross-Hellauer *et al.*, 2022).

2.4. Infrastructuring in Science and Technology Studies (STS) and the Challenge of Trust and Credibility in Research Infrastructures (RIs)

Infrastructure has garnered renewed interest from STS scholars since the 2010s. One of the guiding questions of this revival has been how to address diverse infrastructures (e.g. transportation, information infrastructures, electricity, etc.) under a single rubric despite their heterogeneity. This refocusing has been called 'infrastructure inversion' (Star and Ruhleder, 1994) to emphasize that infrastructure is not a neutral background, but has political consequences: modern nation states could not have coalesced without the expansion of print media (Reicher, 2013). Slota and Bowker (2017, p. 531) remark that 'one of the key insights of STS has been to treat infrastructure relationally: it is not so much a single thing as a bundle of heterogeneous things (standards, technological objects, administrative procedures – in Foucault's term, a *dispositif technique* (Foucault, 1979) – which involves both organizational work as well as technology', and further:

> 'The centrality of the material in anthropological infrastructure studies engenders a discussion of 'embodied experience governed by the ways infrastructures produce the ambient conditions of everyday life: our sense of temperature, speed, fl[u]orescence, and the ideas we have associated with these conditions' (Larkin 2013, 335). Interactions with infrastructure govern not just the aesthetic experience of the world, they define imaginaries of what is possible and potentially possible and are presented politically as a pathway to those potentials.' (Slota and Bowker, 2017, p. 535)

STS scholars have increasingly been using the retronym 'built infrastructure' to describe what infrastructures used to be before the advent of information infrastructures. Similiarly, David (1990) found a cultural lag between innovation and usability (Slota and Bowker 2017, p. 536), before the invention of the appropriate 'infrastructural imaginary'. For Research Infrastructures, this would be the function of the Open Science discourse, respectively the FAIR Principles (Wilkinson *et al.*, 2016). Further, many have pointed to the network effects of information/communication infrastructures (where additional users do not increase individual costs associated with using a network, but entail additional benefits).

With 'The Ethnography of Infrastructure', Susan Leigh Star (1999) presents a methodological treatise on how to study infrastructure. According to Star, infrastructure is both ecological and relational, part of actions, tools, and the built environment. The ecology of the distributed high-tech workplace is impacted by infrastructure that permeates all its functions. In order to fully appreciate how an infrastructure works, one needs to examine those who are *not* served by it, and to examine technological systems in the making – what counts as infrastructure is a matter of perspective. Infrastructure is

fundamentally relational (Star and Ruhleder 1996, p. 113) and encodes work (Star 1999, p. 385), from which follows the methodological imperative to find the invisible work needed to sustain a given infrastructure (cf. Leonelli, 2016).

Infrastructure tends to mean different things to different groups. Star and Ruhleder were arguably the first to formalize many of the preceding concepts of infrastructure, influencing future methodological work in infrastructure studies. Their 'Steps towards an Ecology of Infrastructure' (1996) was an ethnographic study of a collaboratory of biologists and computer scientists working with the Worm Community System (WCS), a digitized library of C. elegans flatworm specimens and technologically mediated paths for collaboration among the biologists working with them, which was in many ways an ideal site for the implementation of new computing infrastructure (social expectation of collaboration and a well-established network of biologists sharing specimens). WCS was a failed project with little uptake among the studied communities. In their accounting for that failure, Star and Ruhleder propose infrastructural issues as major factors influencing that outcome (Slota and Bowker, 2017, p. 537): 'As Star and Ruhleder (1996) have argued so eloquently, one person's invisible infrastructure is another person's job, to be faced materially and directly every day. Infrastructure, as they argue, is inherently relational – a given system, technology, or organization is infrastructural to a particular activity at a particular time' (Slota and Bowker 2017, p. 531).

The upshot of this discussion is as follows. Infrastructure has multiple dimensions: it is both embedded and transparent; it exists (metaphorically) underneath other social, technical (built) worlds and does not need to be reconsidered every time it is 'used' to enable a task. It tends to become visible only upon breakdown; it embodies standards and practices that are learned as part of enculturation processes into a given user community; it is rarely built de novo and tends to be fixed in modular increments (Star, 1999, p. 381 ff.).

Infrastructure is inherently technical, social, organisational, political, as well as (in the case of research infrastructures (RIs)), epistemic (Edwards, 2010). STS applies sociotechnical concepts to RIs (Slota and Bowker, 2017, p. 537): 'For Hughes and Latour, infrastructure was not just technology: it was always already braided with social, cultural, and political actors and their values' (Slota and Bowker, 2017, p. 532). Actor-Network Theory was developed largely in an ethnomethodological tradition, but with the important addition of stressing the interchangeability of human and non-human actors (Latour, 1993): Technology is politics by other means (Latour, 1987). There has been a return to physical infrastructures in STS after 2010, with increased interest in the material aspects of knowledge production (Edwards, 2010), even before the advent of the Open Science movement and its renewed interest in knowledge infrastructures and research practices (across STS, CSCW, Information Science, etc.). Another important strand of literature discusses information infrastructures and changes to the profession of the librarian over the last four decades. The advent of Open Science, with the establishment of data

professionals (data stewards) changes the profession yet again. Since then, STS developed an interest in the epistemic affordances of infrastructures. Edwards (2010) details how climate is established as a global phenomenon through the construction of a global climate observational infrastructure. With the dominance of data, STS focuses increasingly on the dominance of computer and data science over 'domain science' (Ribes, 2017; Ribes *et al.*, 2019). The situation resembles the displacement of earlier infrastructures (e.g. Xerox machines) in a process of colonisation which tends to remodel the colonised practices in its own image. EOSC as a federation of RIs seems particularly amenable to STS discussions of infrastructure via ANT's contention that agency is distributed between humans and non-humans (Latour, 1987). Latour and Woolgar (1986) had already drawn attention to the material substrates of intellectual life ('immutable mobiles'), as had Knorr Cetina (1981). Edwards *et al.* (2013) postulate a continuity from the study of physical infrastructure to the study of cyberinfrastructures (e.g., in terms of path dependence). The need for standardization (Busch, 2011) tends to create problems of its own, e.g. in defining a 'standard human', or in defining a 'standard research process'). In this, STS work has also had its 'infrastructural inversion', in beginning to describe the history of large-scale systems as part of human organization (Edwards, 2010).

3 Research Design & Methodology

The research described below was conceived against the backdrop of the EOSC Focus project[3] and activities of the EOSC Support Office Austria[4] Working Group on Researcher Engagement in Austria. Both initiatives are geared towards shaping the European Open Science Cloud (EOSC). A survey as described in section 3.1 was launched as part of EOSC Focus, while the WG conducted semi-structured interviews as outlined in section 3.2. The survey was used to enable quantitative analysis, offering the typical advantages associated with standardized surveys, such as broad reach, anonymization, efficiency in data collection and analysis, and relatively low demands on personnel resources. In parallel, semi-structured interviews allowed participants to raise relevant issues that may not have been anticipated in the survey. The design and implementation of both the survey and interviews are described below concluding with limitations of the research design in section 3.3.

[3] https://eosc.eu/eosc-focus-project/

[4] https://eosc-austria.at/

3.1 Survey Design and Methodology

The survey was sent to an entire cohort of students of Data Science (N=120, response rate=75%). 44,33% of participants were also working professionally in the field, with an average of 2.86 years of experience. It was designed to understand factors affecting trust as well as quality indicators used by respondents with respect to sharing and reusing open-source software, data, tools and services. The survey consisted of 35 open and closed questions, organised into six sections. Section 1 (4 questions) collected information on the participants' practical experience. Section 2 (12 questions) addressed respondents' code-sharing practices. Section 3 (5 questions) focused on code reuse, and section 4 (7 questions) explored aspects of quality. Trust was addressed in section 5 (5 questions). The survey concluded with a question on accountability, and another asking for additional comments. The anonymised data set[5] and the survey[6] are available on Zenodo, while the interim results were presented and discussed at the 18th edition of the International Digital Curation Conference 2024 (IDCC24) and published (Flicker *et al.*, 2024).

The survey was circulated via TUWEL, TU Wien's e-learning platform. The standardized questions were thus analysed automatically; the open-ended questions were analysed using inductive categorisation to derive categories ('codes') from the text responses. These 'codes' are intended to represent the material and thus allow statements and interpretations without distorting the core content of the material (Mayring, 2013). The qualitative analysis was supported by the software ATLAS.ti (Kelle, 2013).

3.2 Semi-structured interviews

Additionally, twelve semi-structured interviews were conducted by the EOSC Support Office Austria's (EOSC SOA) Working Group Researcher Engagement in Austria (WG REA). The interview guide was developed iteratively by all members of the WG REA, and structured in three parts: *Part 1* collected interviewees' demographic information (disciplines, institutional and departmental affiliation, seniority level, position and career stage, and gender). Two introductory questions related to the interview partners' research and their motivations aiming at more information to contextualize all information given on research and data practices, trust and data quality. *Part 2* focused on actual data practices. *Part 3* dealt with trust in data and data quality. The interview guideline is available on Zenodo[7].

[5] https://zenodo.org/records/11176945

[6] https://zenodo.org/records/10626345

[7] https://zenodo.org/records/15295668

Interview participants were recruited from Austrian public universities, to ensure the collection of nationally relevant perspectives and requirements, and to accommodate for the limited resources available to the WG members[8].

The participants were recruited through personal contacts, additional researchers were approached via email using a purposive, though not fully systematic, strategy, with a particular focus on department heads and recipients of ERC grants, individuals who are typically well-established within their disciplines and thus possess substantial familiarity with research infrastructures, environments, and practices. Furthermore, it could be expected that their professional standing and credibility within the research community would increase the likelihood that their views were influential at both peer and institutional levels. These individuals were also well positioned to disseminate the interview request within their teams. Respondents were recruited from fields as diverse as Computer Sciences, Life Sciences, Social Sciences and Humanities, covering ten Austrian public universities (out of 23). The interviews were conducted, recorded and transcribed by five members of the WG REA. The transcripts were later edited for legibility.

Para-linguistic features of spoken language (laughter, sighs or breathing) or extra-linguistic features (eye movements, gestures) were not transcribed, while word repetitions and interruptions or slips of the tongue were written down (Kowal and O'Connell, 2013).

Interviews were edited for publication on Zenodo[9], in consultation with the interviewees. It is crucial to note, that the transcribed interviews – not the interviews edited for publication – were used for the analysis. Interview analysis followed the inductive categorization approach (Mayring, 2013) and was supported by ATLAS.ti (Kelle, 2013). After a first round of analysis, applied codes as well as interim results were discussed in small groups.

3.3 Limitations of the research design

Four factors limit the interpretability of the results. First, the original purpose of the WG REA was not to improve the understanding of research practices, nor to develop indicators that facilitate trust in data and data quality, services and tools, but to collect requirements for research environments and infrastructures from Austrian researchers, and feed them into strategic (inter-) national policy papers to inform the development of EOSC. Consequently, the study prioritized research institutions located in Austria, while

[8] This approach was also chosen for pragmatic reasons: public universities are officially recognized research institutions and are listed by oesterreich.gv.at, a platform across public authorities. The Austrian Federal Chancellery is responsible for the content of this platform.

[9] https://zenodo.org/communities/wgreaeoscsoa/records?q=&l=list&p=1&s=10&sort=newest

simultaneously striving for an inclusive survey approach that encompassed a broad range of scientific disciplines.

Second, the WG REA members come from diverse professional backgrounds, including the library sector, social sciences, data science, physics, and business informatics. As a result, their familiarity with interviewing as a method of data collection and transcription and as a theoretically informed - but not neutral - practice varied considerably. While many issues relating to transcription were addressed through group discussions (e.g., regarding what should be transcribed), the diversity in experience contributed to inconsistent interviewing styles and occasional errors during the interviews (Hopf, 2023). Despite all interviewers being well-acquainted with the project, challenges remained in determining when to steer conversations back to the central topic or when to probe further with follow-up questions. These difficulties were compounded by disciplinary differences: interviewers often came from fields different from those of the interviewees, which sometimes hindered their ability to recognise opportunities for content-relevant follow-up questions. In several cases, such gaps were only identified retrospectively during transcript review and subsequent research.

Third, due to limited time and financial resources, only a single round of analysis was conducted for the open-ended survey responses and interviews. Coding and results were discussed, reviewed, and revised within small groups, but not by the entire research team. As such, the findings should be interpreted with appropriate critical consideration.

Last, regarding representativeness of the survey, it is important to note the implications of the sampling strategy and data collection procedure as described in ‘3.1 Survey Design and Methodology’. The findings are generalizable to the entire cohort of Data Science students at TU Wien but do not necessarily constitute a representative sample of data scientists more broadly. A more fundamental limitation of these responses concerns the fact that these likely represent ideal (as opposed to actual) data practices, in the same way that research papers model ideal, not actual, laboratory practices (Knorr Cetina, 1981).

4 Results and Discussion

This section presents and discusses the results jointly. Instead of separating findings by how the data was collected (the survey and the semi-structured interview series) the chapter is organized into three thematic sections reflecting the key findings: Perceiving Trustworthiness (4.1), Indicators of Quality (4.2), and Documentation (4.3). This structure was chosen because insights from both the survey and the interviews contributed to each of these themes.

4.1 Perceiving Trustworthiness

In the context of scientific knowledge production, trust needs to be rational as well as evidence based. In other words, there needs to be a reason grounded in evidence that someone or something is trustworthy. Reasoning must be revised in case new evidence shows up (McLeod, 2021). It is thus linked to quality management, or rather to methods to check quality that differ depending on the actual discipline and related research questions.

Additionally, it is – and that across disciplines – also linked to context, or more specifically to organisations, institutions, and (peer-reviewed) journals, although the latter has been criticised for being flawed (lack of time, highly subjective perspectives). In other words, reputation matters. This leads to further questions about how organisations, institutions and journals can both build and maintain a reputation for being trustworthy, or about how reputation can be communicated in dynamic environments or in case potential recipients come from outside the core community or different geographic regions. While there is currently no answer to the second question, there are certainly hints on how reputation can be maintained in a scientific context. As one participant in the semi-structured interview series noted: ‘In my opinion, trust in this context is only possible if I can assume that my colleagues have the same medical ethics or scientific ethics.’ (Schmutzhard and Flicker, 2023)

4.2 Indicators of Quality

Although previous studies have partially explored the discrepancies between ideal research practices (Flicker *et al.,* 2024) or the motivations for data sharing and reuse (Reichmann *et al.*, 2021) and actual research conduct, the combination of the survey and the interview series offers novel insights: In the context of sharing data, tools, services and open-source software, the survey suggested a strong discrepancy between the ideal of quality checks and their implementation. While many stated that quality checks should be conducted before re-use to determine fitness for purpose, they failed to live up to that (mostly due to lack of time). Rather, they decided to trust whatever they intended to reuse based on specific indicators.

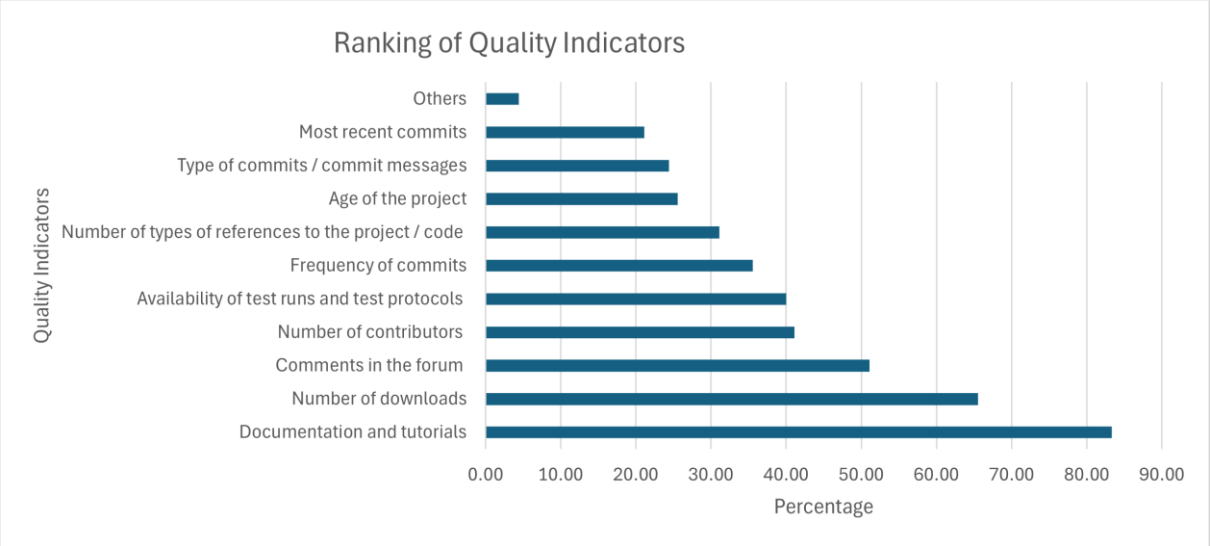


Figure 1: Quality Indicators as ranked by students and professionals of Data Science.

Fig. 1 shows the quality indicators believed to be most crucial based on the survey's quantitative analysis. The top ranked quality indicator was 'Documentation', followed by the 'Number of downloads' and 'Comments in the Forum'. While 83.33% considered 'Documentation and tutorials' to be the most important, the 'Number of downloads' was almost 20% lower at 65.56%. Other quality indicators were 'Number of contributors' (41.11%), 'Availability of test runs and test protocols' (40%), 'Frequency of commits' (35.56%), 'Number of types of references to the project / code' (31.11%) and 'Age of the project' (25.56%), 'Type of commits / commit messages' (24.44%) and 'Most recent commits' (21.11%). The final category was 'others' (4.44%).

The qualitative analysis of the survey draws a similar, yet somewhat different picture as is shown in the Sankey Diagram in **Fig. 2**. The left side lists the quality indicators: (i) (Active) Community, (ii) Documentation, (iii) Reproducibility, (iv) Comments, (v) Commits, (vi) Downloads, (vii) Popularity, (viii) Test protocols, (ix) Tutorials, (x) Availability of test runs and (xi) References. The wider a bar, the more frequently an indicator was mentioned.

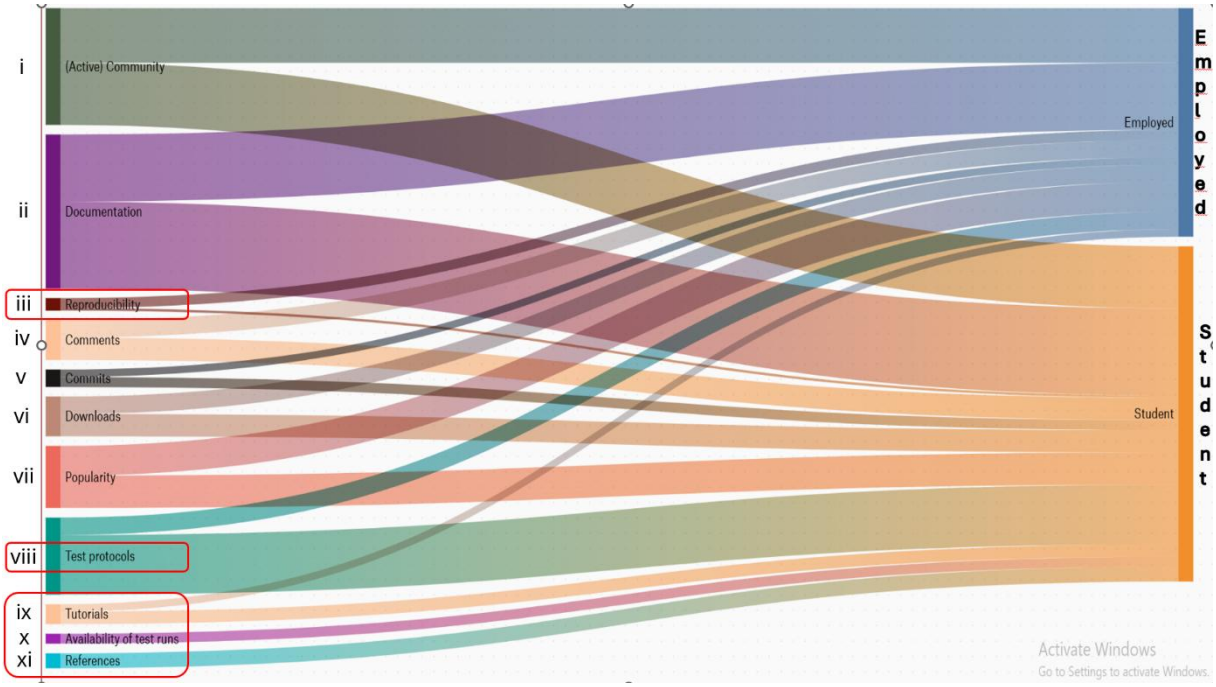


Figure 2: Sankey Diagram showing the frequency with which quality indicators ((Active) Community (i), Documentation (ii), Reproducibility (iii), Comments (iv), Commits (v), Downloads (vi), Popularity (vii), Test protocols (viii), Tutorials (ix), Availability of test runs (x) and References (xi)) are cited by two groups – namely those who are Employed and Students. Indicators (iii, viii, ix, x and xi) with differing ratings between the two groups are highlighted in red.

The right side separates the survey's participants into two groups - Students and those who were also employed in subject-related fields (Employed)[10]. The stream-shaped connections between the indicators on the left and the groups of respondents on the right provide an approximate visualization of which groups identified specific indicators and the frequency with which they did so.

The most frequently cited quality indicators are (i) (Active) Community, (ii) Documentation, (vii) Popularity, and (viii) Test Protocols. The analysis of the stream-like connections reveals that (i) (Active) Community, (ii) Documentation, and (viii) Popularity are considered important by both groups. However, (i) (Active) Community and (vii) Popularity appear to hold slightly greater significance for students. (viii) Test Protocols are predominantly valued by students, whereas (iii) Reproducibility is more relevant to Data Scientists. Two indicators – (x) Availability of Test Runs and (xi) References – were mentioned exclusively by students. In other words, certain quality indicators are exclusively, or at least predominantly, relevant to students ((viii)Test Protocols, (ix) Tutorials, (x) Availability of test runs, (xi) References), whereas others are more applicable to professionals ((iii) Reproducibility).

[10] These results as well as the graphic are based on interim findings after the analysis of around one third of the survey.

4.3 Documentation

Documentation (ii), emerges as the most important quality criterion for both groups, as confirmed by two independent methods of analysis. This finding is further supported by the results of the interview analysis. For example, one Computer Scientist from the semi-structured interview series put it that way:

> ‘We collect provenance information, i.e., how the data is being defined, being generated, being collected and so on and so forth, to provide at least some basic information about the Quality and what can be expected out of them. Documenting the original goals of data generation helps to ensure that datasets are not taken out of context and used for things that simply do not align with them at all.’ (Ekaputra and Czuray, 2024)

A Communication Scientist who participated in the same interview series emphasised the relevance of documentation when stating that ‘Qualitative Research (...) is enormously context sensitive. In order to use them [data] again, it would probably require a great deal of contextual information and metadata, since, for example, it would have to be known who created, processed, analysed and interpreted the data, when, how and why and under what limitations.’ (Schreiber and Flicker, 2023), while a linguist interviewed in the same context supported the need to contextualize research data via metadata:

> ‘I think it is crucial to contextualise the data, which is very important in cultural studies. That usually happens through the metadata, but in the case of social media that might not be enough.’ (Reichl and Blumesberger, 2024)

Closer examination during the interviews, and further analysis of the survey data, revealed that responses to the question of what should be documented, or what constitutes good documentation, were often superficial and/or highly variable. Among other aspects, documentation was associated with elements such as metadata, the provision of information on prior use and its outcomes, illustrative use cases and examples, as well as test protocols and test data. While this may partially reflect limitations in the data collection design, the variability underscores the need for further investigation given the importance of documentation as a quality indicator.

Furthermore, it is important to recognise that quality is not defined by universally applicable criteria; rather, it is context-dependent and should be evaluated in terms of fitness for purpose within specific research settings. The indicators discussed above serve as tools to assess whether a given resource can and should be reused for scientific purposes within research contexts.

Considering the implications for the design and development of research infrastructures is therefore important. For instance and according to findings of the semi-structured interview-series, peer-reviewed journal publications are often regarded as more trustworthy than pre-print versions:

> ‘I am not a fan of data that has appeared in the context of pre-prints – although there seems to be a trend towards this. Many of these pre-prints unfortunately get stuck at exactly this stage and are never published in a recognised journal. This already raises the question of what is wrong with these publications and how reliable are data from these articles that have not been accepted by reviewers.’ (Hofer and Flicker, 2023)

However, the peer review process has also been subject to criticism, particularly regarding the unpaid and time-constrained contributions of researchers, and the perceived subjectivity and lack of transparency in evaluations. These concerns highlight the need to explore practical improvements to the peer review system.

More broadly, it is important to consider how infrastructure-integrated processes can be designed and implemented to promote not only the quality, trust, and trustworthiness of research, but also of associated research artifacts such as data and code. As potential initial steps, researchers highlighted the importance of establishing rules and regulations governing infrastructure use – for example, mandatory standards for data submission to repositories – bringing up questions about how compliance with standards would be monitored and what consequences would follow in cases of non-compliance. While these issues fell outside the scope of the present study, they warrant further investigation.

5 Conclusions and Future Work

The results presented so far form a solid basis for informing and shaping the design of research infrastructures as well as for follow-up studies to elicit concrete guidelines on mechanisms, processes, and human factors increasing both trust and trustworthiness when developing, deploying and operating data services and research infrastructures. Follow-up studies, however, are in need of focus in terms of both disciplines and topics. In addition to the necessity of concentrating on a limited number of disciplines, the survey and interview series have identified several potential avenues for future research. These include quality management and quality checks, the relevance of how reputation facilitates trust and questions about how to not only design but also implement infrastructure-integrated processes promoting quality, trust and trustworthiness regarding research and its artifacts such as data and code. Regarding the latter, the importance of transparent rules and regulations such as mandatory standards for data uploading was emphasized, highlighting key governance-related concerns. Further research in this context would also necessitate an investigation into the mechanisms for monitoring compliance and consequences in cases of non-compliance.

Researchers noted that reputation becomes a critical factor when the processes of checking and verifying data, services, and tools for reuse become exceedingly difficult, time-consuming, or even unfeasible. In such cases, the trustworthiness of the source – such as research organizations, repositories, or journals – gains importance. One factor contributing to perceived trustworthiness is the extent to which these sources adhere to

established scientific principles and ethical codes. Further research is needed to identify additional indicators of trustworthiness and to explore how reputation can be effectively communicated in dynamic research environments, particularly when stakeholders originate from outside specific communities of practice or from different geographic regions.

In the context of quality management and quality assurance, it is important to recognize that, given the diversity of research practices and objectives, the concept of ‘fitness for purpose’ – that is, suitability for a specific research question or context – may be more appropriate than the notion of general quality.

Documentation is key for assessing fitness for purpose. However, how documentation is perceived and evaluated varies across social settings warranting further investigation. Additional fitness for purpose criteria should be identified, and research should explore the extent to which their development and implementation can be automated or integrated into existing research infrastructures. Such efforts must be mindful of community acceptance to avoid proposing standards that are impractical or unlikely to be adopted. Furthermore, the robustness of these criteria against intentional manipulation should also be critically examined.

A discrepancy between research ideals and practical realities was also identified – the expectation that data, services, and tools intended for reuse should be evaluated for their fitness for purpose prior to reuse, and the practical challenges (including time constraints) preventing consistent implementation of this requirement. Thus, further questions emerge regarding the identification of indicators that can support the assessment of fitness for purpose and concerning mechanisms enabling researchers to maintain confidence in their work and assume liability, even when full verification is not feasible.

Acknowledgements

This research was partially funded by the EOSC Focus project (Horizon Europe, Grant No. 101058432). The views expressed are those of the authors and do not reflect those of the European Union.

References

Barber, Bernard. (1983) *The Logic and Limits of Trust*. New Brunswick, NJ: Rutgers Univ. Press.

Begley, C. G. and Ioannidis, J. P. A. (2015) 'Reproducibility in Science: Improving the Standard for Basic and Preclinical Research', *Circulation Research*, 116(1). doi: https://doi.org/10.1161/CIRCRESAHA.114.303819

Boeckhout, M. et al. (2018) 'The FAIR Guiding Principles for Data Stewardship: Fair Enough?', European Journal of Human Genetics, 26(7): pp. 931–36. doi: https://doi.org/10.1038/s41431-018-0160-0

Borgman, C. L. and Groth, P. (2025) 'From Data Creator to Data Reuser: Distance Matters', Harvard Data Science Review, 7(2). doi: https://doi.org/10.1162/99608f92.35d32cfc

Busch, L. (2011) *Standards: Recipes for Reality*. Cambridge, Massachusetts: The MIT Press. doi: https://doi.org/10.7551/mitpress/8962.001.0001

Collins, H. (2007) *Rethinking Expertise*. Chicago, Ill.: University of Chicago Press.

Daniel, H. D. and Mittag, S. and Bornmann, L. (2007) 'The Potential and Problems of Peer Evaluation in Higher Education and Research', *Quality Assessment for Higher Education in Europe*, pp. 71–82.

David, P. (1990) 'The Dynamo and the Computer: An Historical Perspective on the Modern Productivity Paradox', *American Economic Review*, 80(2), pp. 355–61.

DiPrete, T. A. and Eirich, G. M. (2006) 'Cumulative Advantage as a Mechanism for Inequality: A Review of Theoretical and Empirical Developments', *Annual Review of Sociology*, 32(1), pp. 271–97. doi: https://doi.org/10.1146/annurev.soc.32.061604.123127

Edwards, P. et al. (2013) *Knowledge Infrastructures: Intellectual Frameworks and Research Challenges*. Available at: https://knowledgeinfrastructures.org/2012-workshop-report/ (Accessed: 3 June 2025)

Edwards, P. N. (2010) *A Vast Machine. Computer Models, Climate Data, and the Politics of Global Warming*. London, Cambrige, MA: The MIT Press.

Ekaputra, F. J. and Czuray, M. (2024) 'EOSC Support Office Austria: Visions, needs and requirements for research data and practices. An interview with Fajar J. Ekaputra'. Available at: https://zenodo.org/records/14332331

Flicker, K. *et al*. (2024) 'Factors influencing Perceptions of Trust in Data Infrastructures, *International Journal of Digital Curation*, 18(1), pp. 1-12. doi: https://doi.org/10.2218/ijdc.v18i1.921

França, T. F. A. and Monserrat, J. M. (2019) 'Reproducibility Crisis, the Scientific Method, and the Quality of Published Studies: Untangling the Knot', Learned Publishing, 32(4), pp. 406–8. doi: https://doi.org/10.1002/leap.1250

Gambetta, D. (1988) 'Can We Trust Trust?', In Gambetta, D. (Ed.) *Trust: Making and Breaking Cooperative Relations*, New York u.a.: Blackwell, pp. 213–38.

Giddens, A. (1996) *The Consequences of Modernity*. Cambridge: Polity Press.

Hardin, R. (2001) 'Conceptions and Explanations of Trust'. In Cook, Karen S. (Ed.) *Trust in Society*. New York: Russell Sage Foundation, pp. 3–39.

Hardin, R. (2006) *Trust*. Cambridge, UK: Polity.

Hawley, K. (2014) 'Trust, Distrust, and Commitment', *Noûs* 48(1), pp. 1-20. doi: https://doi.org/10.1111/nous.12000

Hofer, T. and Flicker, K. (2023) 'EOSC Support Office Austria: Visions, needs and requirements for research data and practices. An interview with Thomas Hofer'. Available at: https://zenodo.org/records/7855137

Hopf, C. (2013) 'Qualitative Interviews - ein Überblick' in Flick, U., von Kardorff, E. and Steinke, I. (eds.) *Qualitative Forschung. Ein Handbuch*. Leipzig: Rowohlt Taschenbuch Verlag, pp. 349-360.

Horbach, S. P. J. M. and Halffman, W. (2018) 'The Changing Forms and Expectations of Peer Review', *Research Integrity and Peer Review*, 3(1), pp. 8. doi: https://doi.org/10.1186/s41073-018-0051-5

Hosseini, M. *et al.* (2024) 'Messing with Merton: The Intersection between Open Science Practices and Mertonian Values', *Accountability in Research*, 31(5), pp. 428–55. doi: https://doi.org/10.1080/08989621.2022.2141625

Ioannidis, J. P. A. (2005) 'Why Most Published Research Findings Are False', PLOS Medicine, 2(8), e124. doi: https://doi.org/10.1371/journal.pmed.0020124

Jacovi, A. et al. (2021) 'Formalizing trust in Artificial Intelligence: prerequisites, causes and goals of human trust in AI'. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 624–635. doi: https://doi.org/10.1145/3442188.3445923

Kelle, U. (2013) 'Computergestützte Analyse qualitativer Daten' in Flick, U., von Kardorff, E. and Steinke, I. (eds.) *Qualitative Forschung. Ein Handbuch*. Leipzig: Rowohlt Taschenbuch Verlag, pp. 485-502.

Knorr Cetina, K. (1981) *The Manufacture of Knowledge. An Essay on the Constructivist and Contextual Nature of Science*. Oxford: Pergamon Press.

Kowal, S. and O'Connell, D. C. (2013) 'Zur Transkription von Gesprächen' in Flick, U., von Kardorff, E. and Steinke, I. (eds.) *Qualitative Forschung. Ein Handbuch*. Leipzig: Rowohlt Taschenbuch Verlag, pp. 437-447.

Latour, B. (1987) *Science in Action. How to Follow Scientists and Engineers through Society.* Cambridge, MA: Harvard University Press.

Latour, B. (1993) *The Pasteurization of France.* Cambridge, MA: Harvard University Press.

Latour, B. and Woolgar, S. (1986) *Laboratory Life. The Construction of Scientific Facts*. Princeton, NJ: Princeton University Press.

Leonelli, S. (2016) *Data-Centric Biology. A Philosophical Study*. Chicago, London: The University of Chicago Press.

Leonelli, S. (2018) 'Re-Thinking Reproducibility as a Criterion for Research Quality', *Including a Symposium on Mary Morgan: Curiosity, Imagination, and Surprise (Research in the History of Economic Thought and Methodology)*, 40A.

Leonelli, S. (2020) 'Learning from Data Journeys', in Leonelli, S. and Tempini, N. (eds.) *Data Journeys in the Sciences*. Cham: Springer International Publishing, pp. 1-12. doi: https://doi.org/10.1007/978-3-030-37177-7_1

Leonelli, S. and Tempini, N. (2020) Data Journeys in the Sciences. Cham: Springer International Publishing. doi: https://doi.org/10.1007/978-3-030-37177-7

Lewis, J. D. and Weigert, A. J. (1985) 'Trust as a Social Reality', Social Forces, 63(4), pp. 967–85. doi: https://doi.org/10.1093/sf/63.4.967

Lewis, J. D., and Weigert, A. J. (2012) 'The Social Dynamics of Trust: Theoretical and Empirical Research, 1985-20121', Social Forces, 91(1), pp. 25–31. doi: https://doi.org/10.1093/sf/sos116

Luhmann, N. (1979) *Trust and Power. Two Works by Niklas Luhmann. With Introduction by Gianfranco Poggi.* Chichester: Wiley.

Mayring, P. (2013) 'Qualitative Inhaltsanalyse' in Flick, U., von Kardorff, E. and Steinke, I. (ed(s).) *Qualitative Forschung. Ein Handbuch*. Leipzig: Rowohlt Taschenbuch Verlag, pp. 468-475.

McLeod, C. (2021) 'Trust'. Available at: https://plato.stanford.edu/entries/trust/ (Accessed: 27.05.2025)

McKnight, D. H. et al. (2011) 'Trust in a specific technology: An investigation of its components and measures', *ACM Transactions on management information systems (TMIS)*, 2(2), pp. 1-25. doi: https://doi.org/10.1145/1985347.1985353

Merton, R. K. (1936) 'The Unanticipated Consequences of Purposive Social Action', *American Sociological Review*, 1(6), pp. 894-904. doi: https://doi.org/10.2307/2084615

Merton, R. K. (1968) 'The Matthew Effect in Science: The Reward and Communication Systems of Science Are Considered', Science, 159(3810), pp. 56–63. doi: https://doi.org/10.1126/science.159.3810.56

Merton, R. K. (1973) 'The Normative Structure of Science', in Storer, N. M. (eds.) *The Sociology of Science*. Chicago: The University of Chicago Press, pp. 267–80.

Nickel, P. J., Franssen, M. and Kroes, P. (2010) 'Can We Make Sense of the Notion of Trustworthy Technology?', Knowledge, Technology & Policy, 23(3), pp. 429–44. doi: https://doi.org/10.1007/s12130-010-9124-6

Oreskes, N. (2019) *Why Trust Science?* Princeton, NJ: Princeton University Press.

Pink, S., Lanzeni, D. and Horst, H. (2018) 'Data anxieties: Finding trust in everyday digital mess', *Big Data & Society*, 5(1). doi: https://doi.org/10.1177/2053951718756685

Reicher, D. (2013) *Nationensport und Mediennation. Zur Transformation von Nation und Nationalismus im Zeitalter elektronischer Massenmedien*. Göttingen, Deutschland: Vandenhoek & Ruprecht.

Reichl, S. and Blumesberger, S. (2024) 'EOSC Support Office Austria: Visions, needs and requirements for research data and practices. An interview with Susanne Reichl'. Available at: https://zenodo.org/records/11357635

Reichmann, S. *et al.* (2021) 'Between Administration and Research: Understanding Data Management Practices in an Institutional Context', Journal of the Association for Information Science and Technology, 11(72), pp. 1415–31. doi: https://doi.org/10.1002/asi.24492

Ribes, D. (2017) 'Notes on the Concept of Data Interoperability: Cases from an Ecology of AIDS Research Infrastructures', in *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. Portland, Oregon, USA: Association for Computing Machinery, pp. 1514–26. doi: https://doi.org/10.1145/2998181.2998344

Ribes, D. *et al.* (2019) 'The Logic of Domains', *Social Studies of Science*, 49(3). doi: https://doi.org/10.1177/0306312719849709

Ross-Hellauer, T. *et al.* (2022) 'Dynamics of Cumulative Advantage and Threats to Equity in Open Science: A Scoping Review', *Royal Society Open Science*, 9(1), 211032. doi: https://doi.org/10.1098/rsos.211032

Schmutzhard, E. and Flicker, K. (2023) 'EOSC Support Office Austria: Visions, needs and requirements for research data and practices. An interview with Erich Schmutzhard'. Available at: https://zenodo.org/records/7868201

Schreiber, M. and Flicker, K. (2023) ‘EOSC Support Office Austria: Visions, needs and requirements for research data and practices. An interview with Maria Schreiber’. Available at: https://zenodo.org/records/7758011

Slota, S. C. and Bowker, G. C. (2017) 'How Infrastructures Matter', in Felt, U. et al. (eds.), *The Handbook of Science and Technology Studies*. Cambridge, Massachusetts: MIT Press, pp. 529-554.

Smith, R. (2006) ‘Peer Review: A Flawed Process at the Heart of Science and Journals’, *Journal of the Royal Society of Medicine*, 99(4), pp. 178–82.

Spiekermann, S. (2016) *Ethical IT Innovation. A Value-Based System Design Approach*. London, New York: CRC Press.

Star, S. L. and Ruhleder, K. (1994) ‘Steps towards an Ecology of Infrastructure: Complex Problems in Design and Access for Large-Scale Collaborative Systems’, in *Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work*. New York, NY, USA: Association for Computing Machinery, pp. 253–64. doi: https://doi.org/10.1145/192844.193021

Star, S. L. and Ruhleder, K. (1996) 'Steps Towards an Ecology of Infrastructure Complex Problems in Design and Access for Large-Scale Collaborative Systems', *Information Systems Research*, 7(1), pp. 111-134. doi: https://doi.org/10.1287/isre.7.1.111

Star, S. L. (1999) 'The Ethnography of Infrastructure', *American Behavioral Scientist*, 43(9), pp. 377-391. doi: https://doi.org/10.1177/00027649921955326

Sztompka, P. (1999) *Trust: A Sociological Theory.* Cambridge u.a.: Cambridge Univ. Press.

Taddeo, M. (2010) ‘Modelling Trust in Artificial Agents, A First Step Toward the Analysis of e-Trust’, Minds and Machines, 20(2), pp. 243–57. doi: https://doi.org/10.1007/s11023-010-9201-3

Taddeo, M. and Floridi, L. (2011) ‘The Case for E-Trust’, Ethics and Information Technology, 13(1). pp 1–3. doi: https://doi.org/10.1007/s10676-010-9263-1

Tronnier, F., Harborth, D., and Hamm, P. (2022) ‘Investigating privacy concerns and trust in the digital Euro in Germany’, *Electronic Commerce Research and Applications*, 53. doi: https://doi.org/10.1016/j.elerap.2022.101158

Walker, M. U. (2006) *Moral Repair: Reconstructing Moral Relations after Wrongdoing*. Cambridge: University Press. doi: https://doi.org/10.1017/CBO9780511618024

Wilkinson, M. D. et al. (2016) ‘The FAIR Guiding Principles for Scientific Data Management and Stewardship’, Scientific Data 3, 160018. doi: https://doi.org/10.1038/sdata.2016.18

Wilkinson, M. et al. (2018). ‘A Design Framework and Exemplar Metrics for FAIRness’, *Scientific Data 5*: 180118. doi: https://doi.org/10.1038/sdata.2018.118

Connecting Feminist STS and Human-Centred Design – a Pathway to Practical Implementation for Practitioners

Charlotte Reinhardt, Nicola Fricke

University of Wuppertal, Germany

DOI 10.3217/978-3-99161-062-5-004, CC BY 4.0

Abstract. Since the 1980s, feminist researchers undertook great efforts to integrate gender considerations into Human-Computer Interaction (HCI) design processes (Ratzer *et al.*, 2021). Despite the significant scholarly contributions made in this area, there remains a notable scarcity of publications that provide concrete, actionable guidance for practitioners. Much of the existing literature tends to focus on broad frameworks or abstract recommendations regarding research attitudes, rather than offering specific guidelines that can be readily applied in practice (Søndergaard, 2018; Chivukula and Gray, 2020; Dankwa and Draude, 2021).

In a recent research, Reinhardt (currently in revision) undertook a comparative analysis of several guidelines that address the integration of gender dimensions into HCI design processes. Through this analysis, we identified four recurring motifs that emerged across the guidelines: 1) a normative design attitude, 2) the body, 3) social constitution and environmental design, and 4) action and interaction. Each of these motifs in turn encompasses several facets that allow for a more differentiated understanding of the implication and ways of application of a truly gender sensitive design.

The primary aim of our long term research is to translate these identified motifs into actionable strategies for practitioners. We assert that the effective application of said motifs within a design process necessitates a participatory approach, engaging diverse stakeholders in the design journey. To facilitate this, in this contribution, we propose a series of reflexive questions tailored to each facet of the identified motifs. These questions are strategically aligned with the four activities outlined in the ISO guideline on human-centred design (HCD) ISO 9241-210. As an intermediate step the reflexive questions serve as the basis for the further development of a practitioner's guideline.

1 Introduction

When designing a technological artefact (TA), designers must consider the characteristics of the target group they are designing for. Past research has shown that the majority of designers tend to develop for a masculine norm, resulting in varying levels of accessibility to TAs for individuals of different genders (Oudshoorn, Rommes and Stienstra, 2004; Rommes, 2014; Offenwanger *et al.*, 2021). In response, several frameworks for gender-inclusive design have been published. Many of these emphasize the participatory aspects of gender-sensitive design, but they do not provide clear guidance on how to include participation in the design process. (Bath, 2009; Rizvi *et al.*, 2022; Stilke and Buchmüller, 2022).

In this article, we present reflexive questions which we will further develop into concrete guidelines for integrating gender sensitive design aspects into an HCD process, as outlined in the ISO standard 9241-210.

To achieve this, we first summarize four recurring motifs in gender-inclusive design identified in a previous study. These motifs are: 1) a normative design attitude, 2) the body, 3) social constitution and environmental design, and 4) action and interaction. We formulate reflexive questions that allow designers to explore different facets of each motif during the design practice.

Next, we provide a brief overview of the HCD design process as described by the ISO standard 9241-210. This standard outlines six design principles and four activities that constitute an iterative design process. We analyse each activity to formulate precise subtasks, thereby providing a more detailed understanding of the design process. Subsequently, we apply the reflexive questions identified earlier to each subtask, offering a first concretisation for integrating the category of sex/gender into an HCD process.

Finally, we will discuss the limitations of this study and provide an outlook on future developments.

2 Recurring motifs in gender inclusive design

Works that focus on the implication of the gender dimension into a Human-Computer Interaction (HCI) design process can be categorized into either gender aware or gender inclusive design approaches (Breslin and Wadhwa, 2018). Being gender aware means that an approach puts its focus on the questions 1) how gender norms, values and behaviour affects the production, use and operation of TAs and 2) how – as a consequence – they are ingrained in those TAs (Breslin and Wadhwa, 2018). Gender inclusive on the other hand are those approaches that do not only seek to find out on the

impact gender has on TAs but also try to change the design accordingly. Gender inclusion '[...] actively seeks to include multiple and intersectional genders, and perhaps even future unknown users and characteristics' (Breslin & Wadhwa, 2018, p. 73). According to the authors within those gender inclusive approaches, two groups can be distinguished: so-called *feminist approaches* and *queer approaches*. Feminist approaches are characterised by activism whereas queer approaches try to overcome the rigidness, that often comes with feminist approaches and try to remain open to actual and potential future users and their usage requirements (Breslin and Wadhwa, 2018).

Vorvoreanu et al. (2019) state that works, that try to contribute to the question how the gender dimension can be implemented into an HCI context, either result in 'demonstration software projects' or in 'methods and practices' (Vorvoreanu *et al.*, 2019, p. 2).

Joining these two classifications this means, that works, that try to implement the sex/gender dimension into the HCI context are either gender aware or gender inclusive. Gender inclusive approaches can be differentiated into either feminist or queer approaches. Each of the three approaches either creates methods and practices or demonstration software.

When designers decide to include the gender dimension into their design, there are a few aspects they need to consider.

Up until today the research on ways to implement the gender dimension into an HCI design process remains rather sparse. As a basis for this article, in 2024 Reinhardt conducted a systematic comparison of three publications that gave direct recommendations on how to implement the gender dimension into an HCI design process. The aim of the systematic comparison was generating a deeper understanding what the term 'gender dimension' means in this context and how it can be operationalised when being 'implemented' into an HCI design process[1].

When it comes to a clear definition of the concept of sex/gender, the vast majority of the studies trying to give recommendations on »gender aware« or »gender inclusive« design, contents with the generally accepted definition of gender as a socially constructed category (Bardzell, 2010; Pollitzer, 2021; Szlavi and Guedes, 2023). Being created as a *critical concept*, the definition of gender as a social category does not offer any answers to the question on how it can be implemented for a productive use.

To obtain a more in depths understanding of the category gender, Reinhardt conducted a hermeneutical analysis (Betti, 1967; Danner, 2006) on publications that give said recommendations to find out, which implications were at work, when gender had been defined as a socially constructed category. Through repeated and comparing lecture of

[1] This comparison is currently in revision.

the three articles, recurring motifs could be extracted. It could be shown, that all works 1) were normatively directed, 2) reflect on the body, 3) reflect on the constitution of the social and the design of the environment and 4) stated the conceptualisation of action and interaction as a central aspect when it comes to the gender dimension in an HCI design process. Each of these motifs fell into several facets, since the articles had enlightened different aspects.

The overarching goal of our research is to offer HCI designers practical guidance on how to include gender in their design processes. As an intermediate step towards achieving this goal, we formulated reflexive questions, that should help designers to derive concrete actions from the motifs for their design process to allow gender sensitive design. Each question corresponds one of the facets described in one of the three articles, that form the basis for Reinhardt's analysis and were clustered in the four motifs.

This contribution is a conceptual work, to link the findings from the feminist science and technology studies to the concept of human centred design. In the next step we will further revise and integrate the reflexive questions together with practitioners for deriving specific guidelines.

The reflexive questions which are mapped to the four motifs are:

A Normative Standpoint

Demand to reflect on the meaning and limitations of one's own standpoint

- What are my convictions? (theoretically and personally?)
- What do I reject? (theoretically and personally?)
- While designing my TA, what do I claim to design and why?

Demand for humility before the user

- Which users or which characteristics of these users do I assume?
- Which characteristics do I subsequently leave out?
- Which physiological concepts do I base my work on?
- Which do I leave out?
- What changes could I make to include a previously excluded group, if that is my aspiration?

Demand to enable to design the environment of a TA

- What influence does the TA have on the environment and is this desirable or not?
- How can people be included into the design process of the design of the environment?

The Body

The body as the subject

- How do I think about the body as a sensing entity?
- What aspects of sensing do I consider in my design?
- What importance do I give to the materiality of my TA?
- Which materials do I choose and why?
- What consequences does this choice of material have for the direct and indirect environment of the TA?
- How does the TA affect the body using it?

The body as the object

- What fundamental image of the human being and the body do I take as the basis for framing the body as the object of a TA?
- What effects does this have on a society?
- Are these effects desirable?

The constitution of the social and the shaping of the environment

The effect on people who use the TA

- What effect do the assumptions I made for this TA have on people and their social environment and is this desirable?
- What roles might be (re)produced by using the TA?

The effect on people who do not use the TA

- In which situations could the TA be used, that has an impact on other people and what does this impact look like and is this desirable?
- What impact does the ecological environment have and is this desirable?

Action and Interaction

Interaction

- In my conception of interaction: who does interact?
- How do I understand interaction between humans and TAs?
- What role does interaction play in my understanding of how society is organized?
- In my opinion, what is the relation between society and interaction?
- What effect does interaction have on the individual?

Action

- Who is acting? (consider human-human and/or human-computer interaction and/or multiple agents' interactions)
- Which aspects of the human being do I consider relevant for action (cognitive processes, body, emotions, etc.)?

These questions were created to enable designers to approach the design of an HCI TA in a gender sensitive way. To ease their use, we suggest to merge them with already existing concepts such as Human-Centred Design.

3 Human-centred design

3.1 What is human-centred design?

The term HCD can be described by numerous definitions, one such definition explains it as 'The process that ensures that the designs match the needs and capabilities of the people for whom they are intended' (Norman, 2013, p. 9). Another one focuses on 'values, concerns, and perceptions of all stakeholders in designing, developing, deploying, and employing products, services, and policies' (Rouse, 2023, p. 4). Based on the ISO 9241-210, the term means designing interactive systems which are user-friendly through using knowledge and techniques from ergonomics and usability research (ISO, 2019). One could shorten the concept to 'honor thy user' (Lee *et al.*, 2017, p. 22) through taking them into account through every step of the design process (Lee *et al.*, 2017).

The term HCD is similar to the term user-centred design (UCD) and often used as a synonym (Billings, 2009) as they both emphasize the human in the centre of research activities. While UCD may include specific methods, such as use-cases in order to derive user-requirements or modelling user-interactions, HCD focusses on understanding user- and other stakeholder-needs (Gasson, 2003). HCD can be understood as a broader concept, maybe even containing UCD. HCD focusses on two aspects: 1) addressing the right problem, and 2) meeting human needs and capabilities (Norman, 2013). One has to keep in mind that the HCD concept developed historically as a counterpart to technology-centred systems design in the tradition of Taylorism (Gill, 1996). Therefore, HCD and UCD both play the same counterpart to technology-centred engineering perspectives.

Norman (2013) specifies four activities as belonging to the HCD process: 1) observation, 2) idea generation, 3) prototyping, 4) testing (Norman, 2013). In his conceptualization, the four phases are organized in spirals with iterative loops in between the phases aiming at an optimization of the created solutions between the iterative loops (Norman, 2013).

The basic ideas of the HCD concept and process formulated by Norman and others have further been formalized within the ISO standard (2019) which we will present followingly.

3.2 The standard 9241-210

The ISO standard 9241-210 (ISO, 2019, p. 6) includes six fundamental principles of HCD:

1) ‘the design is based upon an explicit understanding of users, tasks and environments […];
2) the users are involved throughout design and development […];
3) the design is driven and refined by user-centred evaluation […];
4) the process is iterative […];
5) the design addresses the whole user experience […];
6) the design team includes multidisciplinary skills and perspectives […].’

Alongside these principles, HCD has to be integrated and planned through all phases of the product-development lifecycle (ISO, 2019). The following four design-activities are specified in the ISO (2019, p. 10) standard:

1) ‘understanding and specifying the context of use […];
2) specifying the user requirements […];
3) producing design solutions […];
4) evaluating the design […].’

Since the fourth principle calls for an iterative process, the HCD process can be visualised as the following:

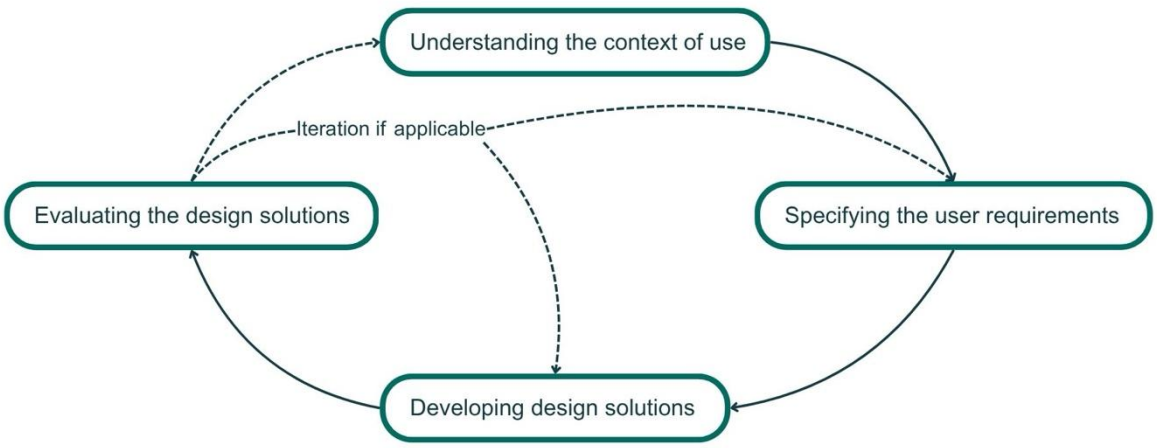


Figure 1.: A simplified visualisation of the HCD process based on the visualisation in: (ISO, 2019).

4 Gendered Design meets Human-Centred Design

4.1 The four activities of human-centred design and their implications

Comparing the findings on gender-sensitive design with the six principles as well as the four activities of HCD, we find structural resemblances.

The motif *A normative standpoint* includes the demand for humility before the user. This humility can be found underlying the principle that demands a fundamental understanding of the user, the task and the context, since the truthful aim to understand somebody is grounded in an openness towards the Other; or as Lee et al. (2017, p. 22) put it, one must 'honor thy user'.

The principle *asking for the user's involvement* resembles heavily the demand to reflect one's own standpoint; as does the principle *asking for design refinement based on user-centred evaluation*. The reflection of one's standpoint is furthermore present in the fifth principle, that demands to involve overall user experience as much as in the sixth principle, that demands to unite interdisciplinary competencies and aspects within the team. Therefore, this first motif works as an addition to the six principles of HCD.

The other four motifs on the other side, rather work as elements, that need to be considered during the four activities of human centred design.

We now would like to integrate the above introduced reflexive questions into the four steps of human centred design. This integration is going to be validated by an expert rating in a future study. In order to do this, it is necessary to generate a deeper understanding of what each activity looks like. We would therefore like to have a closer look on each activity.

Each activity comprehends several subtasks. These subtasks were extracted from the ISO standard. However, they are not identical to the elements formulated in the standard, as they do not only describe subtasks, but also provide general information and therefore operate at different levels. One exception is the activity *evaluation*: The ISO does not foresee subtasks but only provides a differentiation between an inspection-based evaluation and an evaluation with the help of the user. The subtasks were therefore derived analytically.

The first activity according to the ISO standard is to understand and to specify the context of use. Regarding the first principle, this activity asks for nothing less than 'an explicit understanding of users, tasks and environments' (ISO, 2019, p. 6). Hence, analytically the activity *understanding and specifying the context of use* falls into four sub tasks (ISO, 2019):

1.1 a description of users and other stakeholders;
1.2 identification of key characteristics of users or groups of users;
1.3 identification of the user's goals as well as the system's goals;
1.4 identification of the environment of use.

The second activity is to specify user requirements. This activity again falls into four subtasks:

2.1 identify users' and other stakeholders' requirements;
2.2 identify requirements that derive from different aspects (e.g. user context, ergonomics, usability etc.);
2.3 eliminate conflicts between user requirements;
2.4 ensuring the quality of user requirements' specifications.

During the third activity, design solutions are being developed. This falls into five sub tasks:

3.1 designing the task as well as the interaction between the user and the system;
3.2 designing the user interface;
3.3 concretizing design solutions;
3.4 changing the design solutions based on user-centred evaluation and feedback;
3.5 communicating the design solution to the unit responsible for the implementation.

Lastly a user-centred evaluation has to be carried out.

In a gender sensitive design process, analytically this activity needs to happen on different levels, for every aspect of the design process needs to be evaluated to prevent any blind spots:

4.1 Critical investigation of the standard evaluation process;

4.2 Evaluating the design process through a gender sensitive lens;

4.3 Evaluating design solutions.

The critical investigation of the standard evaluation process means, that the evaluation process itself is being looked on.

The standard does not foresee different subtasks, but it provides two different methods on how a user-centred evaluation can take place (ISO, 2019):

A. evaluation with the help of the user and/or
B. inspection-based evaluation.

An evaluation with the help of the user consists of constant involvement of the user in different forms, while an inspection-based evaluation can be conducted with a set of experts and checklists.

Since an inspection-based evaluation that works with checklists implies the existence of checklists, either their development or the choice of existing guidelines can be considered one of the relevant subtasks. The selection of the evaluating experts can be considered a second step within this evaluation method.

On the second level of evaluation, the evaluation offers the opportunity to critically interrogate whether or not the assumptions made during the design process meet the ethical expectations, the designers try to meet. For the evaluation process it is therefore helpful to display all of the assumptions made implicitly or explicitly during the preceding design process. After laying them open, a critical analysis, whether there is room for improvement can take place. It is only then, that design solutions can be evaluated on their own. This step is not specifically gender sensitive; it primarily assesses whether the previously defined design requirements have been fulfilled.

The critical investigation of the standard evaluation process falls into one/two subtasks (depending on the method used):

4.1.A Critical analysis of the user-based evaluation (evaluation with the help of the user);

4.1.B.1Critical analysis of the selection of experts used in the evaluation process (inspection-based evaluation);

4.1.B.2 Critical analysis of the checklists handed out for evaluation (inspection-based evaluation).

The evaluation of the design process through a gender sensitive lens happens through two subtasks:

4.2.1 Disclosure on the explicit and implicit assumptions during the design process;

4.2.2 Critical analysis of the design process through a gender sensitive lens.

The evaluation of the design solutions happens as a result of the preceding evaluation and is an activity on its own.

This way we could identify 18 individual tasks, that we can look at from a gender sensitive perspective.

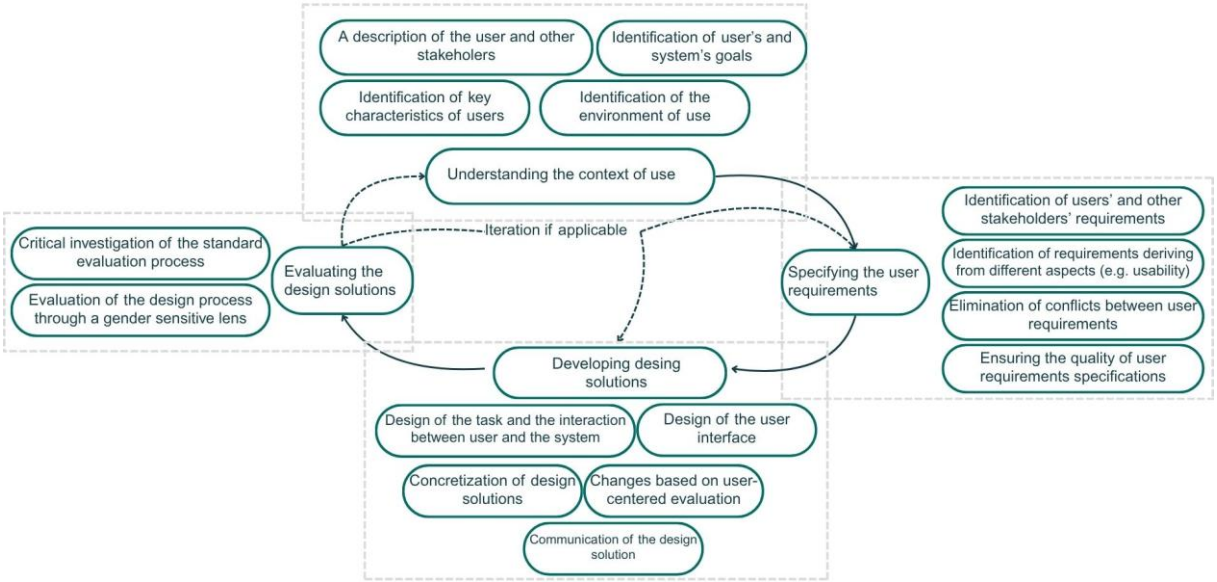


Figure 2.: Illustration of the simplified HCD process including the defined subtasks

4.2 Linking the four activities with the reflexive questions

We are now going through each of the above presented activities and link it to the afore mentioned reflexive questions. To better fit the activities the formulation of the questions has been modified. The reflexive questions can be divided into four different groups of questions: Questions, that invite designers to…

1) reflect on their values and beliefs in a fundamental way;
2) reflect on specific aspects, within a certain subtask;
3) reflect on the ethical implications of their design;
4) critically reevaluate their decisions within a certain subtask.

The reflexive questions were organised according to the four groups, resulting in an order that differs from the sorting introduced above. At this point, it is also relevant to note that not all activities can currently be linked to reflexive questions. Some of the reflexive questions will repeat themselves, however, they are always related to a different object. Lastly, it is necessary to state that the activity *Concretizing design solutions* is identical with the two activities *Designing the user interface* and *Designing the task as well as the interaction between the user and the system.* In order to avoid repetition as far as possible, the questions were not included again.

Understanding the context of use

1. A description of the user and other stakeholders
    a. How do I think about the body as a sensing entity?
    b. Which aspects of the human being do I consider relevant for action (cognitive processes, body, emotions, etc.)?
    c. Which users or which characteristics of these users do I assume? And which do I leave out?
    d. Which body concepts do I base my work on? And which do I leave out?
    e. What aspects of sensing do I consider in my design?
2. Identification of key characteristics of users or groups of users
    a. While designing my TA, what do I claim to design and why?
    b. Why do I consider this specific set of characteristics as key for my design?
    c. What changes could I make to include a previously excluded group, if that is my aspiration?
3. Identification of the user's goal as well as the system's goals
    a. While designing my TA, what do I claim to design and why?
    b. In which situations could the TA be used, that has an impact on other people and what does this impact look like?
    c. What influence does the TA have on the environment and is this desirable?
4. Identification of the environment of use
    a. Which parts of the users' lives do I consider, when thinking about the context of use?
    b. Which physical situations do I consider when thinking about the context of use? (e.g. disabled bodies, pregnancy)
    c. In which situations could the TA be used, that has an impact on other people and what does this impact look like and is this desirable?
    d. What influence does the TA have on the environment? And is this desirable?
    e. Do the aspects, I consider relevant for action, still play the same role in the context of use (e.g. emotional or physical changes)?

Specifying the user requirements

1. Identify users' and other stakeholders' requirements
    a. What is the ideological basis on which I determine the user requirements?
    b. Which users do I assume?
    c. How does the TA affect the body using it?
    d. In which situations, that have an impact on other people, could the TA be used and what does this impact look like?
    e. How can people be included in the design process?
    f. What aspects do I consider in my design and are they pointed out in my participatory design?

g. What importance do I give to the materiality of my TA and is this pointed out in my participatory design?

2. Identify requirements that derive from different aspects
3. Eliminate conflicts between user requirements
    a. In whose favour will the conflict be resolved?
4. Ensuring the quality of user requirement specifications

Developing Design Solutions

1. Designing the task as well as the interaction between the user and the system
    a. How do I understand interaction between humans and TAs?
    b. What role does interaction play in my understanding of how society is organized?
    c. In my opinion, what is the relation between society and interaction?
    d. In my conception of interaction: who does interact?
    e. What effect does interaction have on the individual?
    f. In an HCI situation, who is acting?
    g. What fundamental image of the human being and the body do I take as the basis for framing the body as the object of a TA?
    h. Which aspects of the human being do I consider relevant for action? (cognitive processes, body, emotions, etc.?)
    i. How do I think about the body as a sensing entity?
    j. While designing my TA, what do I claim to design and why?
    k. Which users or which characteristics of these users do I assume? Who or what do I subsequently leave out?
    l. Which body concepts do I base my work on? Which do I subsequently leave out?
    m. What aspects of sensing do I consider in my design?
    n. How does the TA affect the body using it?
    o. How can people be included into designing the environment?
    p. What effects does this have on society and are these effects desirable?
    q. What effect do the assumptions I made for this TA have on people and their social environment and is this desirable?
    r. What roles might be (re)produced by using the TA?
    s. In which situations could the TA be used, that has an impact on other people, what does this impact look like and is it desirable?
    t. What changes could I make to include a previously excluded group, if that is my aspiration?
2. Designing the user interface
    a. How do I think about the body as a sensing entity?
    b. Which aspects of the human being do I consider relevant for action (cognitive processes, body, emotions)?

    c. What aspects of sensing do I consider in my design?
    d. What importance do I give to the materiality of my TA?
    e. Which materials do I choose and why?
    f. What consequences does this choice of material have for the direct and indirect environment of the TA and is this desirable?
    g. How does the TA affect the body using it and is this desirable?
3. Concretizing design solutions[2]
4. Change the design solutions based on user-centred evaluation and feedback
5. Communicate the design solution to the unit that is responsible for the implementation

For the evaluation process it is helpful to display all of the assumptions made implicitly or explicitly during the preceding design process. After laying them open, a critical analysis, whether there is room for improvement can take place.

Critical analysis of the user-based evaluation process

1. Critical analysis of the user-based evaluation
    a. Which body concepts do I base my work on and which ones do I leave out?
    b. Who do I use as evaluator and on what basis was the sample drawn?
    c. How can people be included into the evaluation process?
    d. What changes could I make to include a previously excluded group, if that is my aspiration?
2. Critical analysis of the selection of experts used in the evaluation process (inspection-based evaluation)
    a. In my opinion, what defines an expert?
    b. Which characteristics do I consider relevant for someone to be an expert in this evaluation process?
3. Critical analysis of the checklists handed out for evaluation (inspection-based evaluation)
    a. What does my checklist evaluate?
    b. Who created the checklist?

Evaluating the design process through a gender sensitive lens

Basically, while the disclosure on the explicit and implicit assumptions during the design process displays the assumptions, the second step *evaluating the design process through a gender sensitive lens* asks for the consequences these assumptions have on the world.

[2] See: *Designing the task as well as the interaction between the user and the system* and *designing the user interface.*

1. Disclosure on the explicit and implicit assumptions during the design process
    a. What are my moral, ethical and scientific convictions and what do I reject?
    b. In my conception of interaction: who does interact?
    c. How do I understand interaction between humans and TAs?
    d. What role does interaction play in my understanding of how society is organized?
    e. While designing my TA, what do I claim to design and why?
    f. How do I think about the body as a sensing entity?
    g. What importance do I give to the materiality of my TA?
    h. What fundamental image of the human being and the body do I take as the basis for framing the body as the object of a TA?
    i. In my opinion, what is the relation between society and interaction?
    j. In my opinion, what effect does interaction have on the individual?
    k. Which aspects of the human being do I consider relevant for action (cognitive processes, body, emotions, etc.?)
    l. In a human-computer interaction, who is acting?
    m. Which users or which characteristics of the users do I assume and which do I subsequently leave out?
    n. Which materials do I choose and why?
2. Evaluating the design process through a gender sensitive lens
    a. How can people be included in the design process?
    b. What influence does the TA have on the environment and is this desirable or not?
    c. What consequence does the choice of material have for the direct and indirect environment of the TA and is this desirable?
    d. What effects does the TA have on the body using it and is this desirable?
    e. Are the effects my fundamental image of the human being and the body have on the society desirable or not?
    f. What consequences does my conception of aspects relevant for cognition have on the user and is this desirable?
    g. What effect do the assumptions I made for the TA have on people and their social environment and are they desirable?
    h. What roles might be (re)produced by using the TA and is this desirable?
    i. What impact does the usage of the TA have on other people and is this desirable?
    j. What impact does the TA have on the ecological environment and is this desirable?
    k. What effect does my conception of interaction have on the TA and hence on the society and is this desirable?
    l. What changes could I make to include a previously excluded group, if that is my aspiration?

3. Evaluating the design solutions
    a. What are my convictions (theoretically and personally) and what do I reject?
    b. While designing my TA, what do I claim to design and why?
    c. Which users or which characteristics of these users do I assume? And who or what do I leave out?
    d. What influence does the TA have on the environment? And is this desirable?
    e. What changes could I make to include a previously excluded group, if that is my aspiration?
    f. How can (more) people be included in the design process and is there a better way to do it?

When combining the reflexive questions with the four activities from the standard for HCD, we find, that the dynamic of the process shifts: the fourth activity of the standard had already been evaluation. And since one of the six principles of HCD foresees an iterative process, it can be stated, that evaluation happens several times within an HCD process. Now, that four types of reflexive questions could be identified, of which one once again calls for evaluation, a highly iterative process is created.

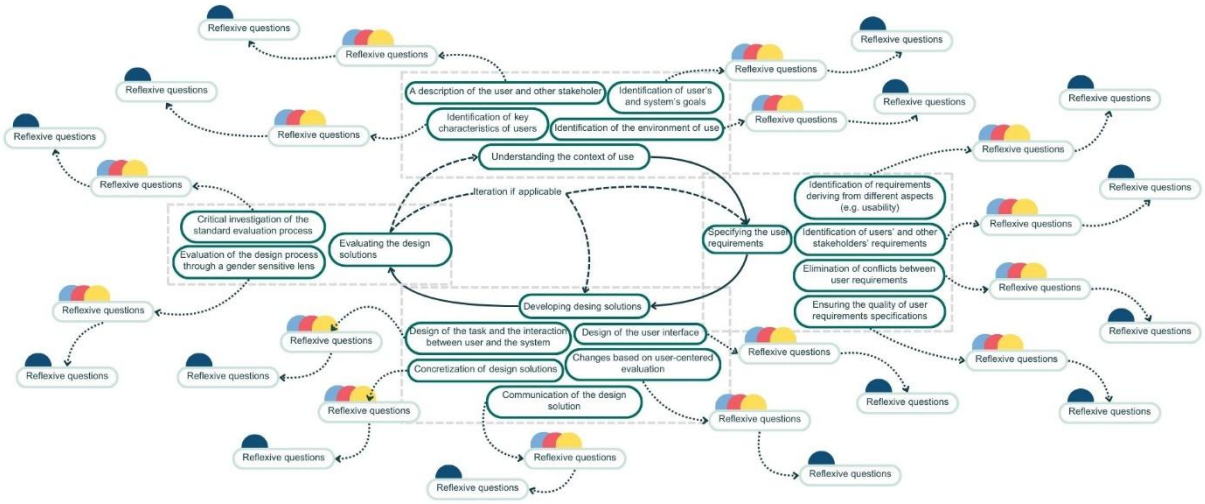


Figure 3.: Illustration of the simplified HCD process including the defined subtasks and the reflexive questions. The Types of questions 1 to 3 are indicated by the colours light blue, red and yellow. The fourth type of questions is indicated by the colour dark blue.

Summary and Outlook

In this article we took a closer look on how to implement the category sex/gender into an HCD process. We started out by describing four recurring motifs, that could be identified in existing frameworks on gender sensitive design and formulating reflexive questions, to elucidate on different facets within each motif. We then provided a short overview of an HCD design process as it is described in the ISO standard 9241-210, focusing on the four design activities, that we analysed to provide a more detailed insight on what each activity includes. We then matched the reflexive questions with the sub tasks to create a first integration of the category sex/gender with an HCD process.

One limitation of this article lies within its sparse literature basis. To further substantiate the data basis, currently we conduct interviews with experts. On this basis we will reevaluate the presented preliminary draft and conduct an expert rating.

This will be further enriched, with general findings from the feminist critique (e.g. on language, power distribution and the link of society and capitalistic structures) as well as findings from adjacent fields, such as feminist architecture, feminist game design, or inclusive HCI design.

Moreover, we will validate the reflexive questions by conducting further expert-interviews and hands-on prototypical usage of the created integration of the questions into the HDC process. In an iterative process the relevant reflexive questions will be modified and adapted to meet developers' needs and for their applicable usage. The outcome will be a shortened list including fewer reflexive questions integrated in the HCD process resembling a guideline for practitioners in order to be more usable in the development process.

References

Bardzell, S. (2010) ‘Feminist HCI: taking stock and outlining an agenda for design,’ in Mynatt, E. D., Hudson, S. E., and Fitzpatrick, G., *CHI 2010: we are HCI : conference proceedings, Atlanta, Ga, USA, April 10-15, 2010*. New York, N.Y.: Association for Computing Machinery, pp. 1301–1310.

Bath, C. (2009) *De-Gendering Informatischer Artefakte: Grundlagen einer kritisch-feministischen Technikgestaltung*. Universität Bremen. Available at: https://media.suub.uni-bremen.de/bitstream/elib/360/1/00102741-1.pdf (Accessed: January 22, 2024).

Betti, E. (1967) *Allgemeine Auslegungslehre als Methodik der Geisteswissenschaften*. Tübingen: Mohr. Available at: https://digitale-objekte.hbz-nrw.de/storage2/2024/01/06/file_3/9484273.pdf.

Billings, C.E. (2009) *Aviation automation: the search for a human-centered approach*. Boca Raton, Fla: CRC Pr (Human factors in transportation).

Breslin, S. and Wadhwa, B. (2018) ‘Gender and Human-Computer Interaction,’ in K.L. Norman and J. Kirakowski (eds.) *The Wiley Handbook of Human Computer Interaction*. 1st ed. Wiley, pp. 71–87. Available at: https://doi.org/10.1002/9781118976005.ch4.

Chivukula, S.S. and Gray, C.M. (2020) ‘Bardzell’s ‘Feminist HCI’ Legacy: Analyzing Citational Patterns,’ in *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. *CHI ’20: CHI Conference on Human Factors in Computing Systems*, Honolulu HI USA: ACM, pp. 1–8. Available at: https://doi.org/10.1145/3334480.3382936.

Dankwa, N.K. and Draude, C. (2021) ‘Setting Diversity at the Core of HCI,’ in M. Antona and C. Stephanidis (eds.) *Universal Access in Human-Computer Interaction. Design Methods and User Experience*. Cham: Springer International Publishing (Lecture Notes in Computer Science), pp. 39–52. Available at: https://doi.org/10.1007/978-3-030-78092-0_3.

Danner, H. (2006) *Methoden geisteswissenschaftlicher Pädagogik: Einführung in Hermeneutik, Phänomenologie und Dialektik*. 5., überarbeitete und erweiterte Auflage. München Basel: Ernst Reinhardt Verlag (UTB Geisteswissenschaften, 947). Available at: https://doi.org/10.36198/9783838509471.

Gasson, S. (2003) ‘Human-Centered vs. User-Centered Approaches to Information System Design,’ *Journal of Information Technology Theory and Application*, 5(2), pp. 29–46.

Gill, K.S. (ed.) (1996) *Human Machine Symbiosis*. London: Springer London. Available at: https://doi.org/10.1007/978-1-4471-3247-9.

ISO (ed.) (2019) 'ISO 9241-210 Ergonomics of human-system interaction – Part 210: Human-centred design for interactive systems.' ISO.

Lee, J.D. *et al.* (2017) *Designing for people: an introduction to human factors engineering*. 3rd edition, revision 1. Charleston, SC: CreateSpace.

Norman, D.A. (2013) *The design of everyday things*. Überarbeitete und erweiterte Auflage. New York, New York: Basic Books.

Offenwanger, A. *et al.* (2021) 'Diagnosing Bias in the Gender Representation of HCI Research Participants: How it Happens and Where We Are,' in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. *CHI '21: CHI Conference on Human Factors in Computing Systems*, Yokohama Japan: ACM, pp. 1–18. Available at: https://doi.org/10.1145/3411764.3445383.

Oudshoorn, N., Rommes, E. and Stienstra, M. (2004) 'Configuring the User as Everybody: Gender and Design Cultures in Information and Communication Technologies,' *Science, Technology, & Human Values*, 29(1), pp. 30–63. Available at: https://doi.org/10.1177/0162243903259190.

Pollitzer, E. (2021) 'Why gender is relevant to materials science and engineering,' *MRS Communications*, 11(5), pp. 656–661. Available at: https://doi.org/10.1557/s43579-021-00093-1.

Ratzer, B. *et al.* (2021) *Enhanced Gender Knowledge and New Content. EECCO. Gender Equality in Engineering through Communication and Commitment.* Projektbericht. Wien: Technische Universität Wien, p. 137. Available at: https://www.tuwien.at/fileadmin/Assets/dienstleister/abteilung_genderkompetenz/gender_in_der_Forschung/GEECCO_Results/Public_deliverables/GEECCO_D6.3_Enhanced_Gender_Knowledge_and_New_Content.pdf (Accessed: May 26, 2025).

Rizvi, N. *et al.* (2022) 'QTBIPOC PD: Exploring the Intersections of Race, Gender, and Sexual Orientation in Participatory Design,' in *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. *CHI '22: CHI Conference on Human Factors in Computing Systems*, New Orleans LA USA: ACM, pp. 1–4. Available at: https://doi.org/10.1145/3491101.3503733.

Rommes, E. (2014) 'Feminist Interventions in the Design Process,' in W. Ernst and I. Horwath (eds.) *Gender in science and technology: interdisciplinary approaches*. Bielefeld: Transcript Verlag (Gender studies), pp. 41–55.

Rouse, W.B. (2023) *From Human-Centered Design to Human-Centered Society: Creatively Balancing Business Innovation and Societal Exploitation*. 1st ed. New York: Productivity Press. Available at: https://doi.org/10.4324/9781003462361.

Søndergaard, M.L.J. (2018) *Staying with the Trouble through Design: Critical-feminist Design of Intimate Technology*. Ph.D. Aarhus University. Available at: https://doi.org/10.7146/aul.289.203.

Stilke, J. and Buchmüller, S. (2022) 'Users and non-users in engineering and feminist participatory research on sustainable aviation,' *NOvation - Critical Studies of Innovation*, (3), p. 110. Available at: https://doi.org/10.5380/nocsi.v0i3.91148.

Szlavi, A. and Guedes, L.S. (2023) 'Gender Inclusive Design in Technology: Case Studies and Guidelines,' in A. Marcus, E. Rosenzweig, and M.M. Soares (eds.) *Design, User Experience, and Usability*. Cham: Springer Nature Switzerland (Lecture Notes in Computer Science), pp. 343–354. Available at: https://doi.org/10.1007/978-3-031-35699-5_25.

Vorvoreanu, M. *et al.* (2019) 'From Gender Biases to Gender-Inclusive Design: An Empirical Investigation,' in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. *CHI '19: CHI Conference on Human Factors in Computing Systems*, Glasgow Scotland Uk: ACM, pp. 1–14. Available at: https://doi.org/10.1145/3290605.3300283.

Longevity Hacking: Ageing as Synthesis in Biomedical Testing

Frederik Peper[1], Nico Wettmann[2]

[1] University of Koblenz, Germany

[2] University of Marburg, Germany

DOI 10.3217/978-3-99161-062-5-005, CC BY 4.0

Abstract. This article examines how ageing becomes visible and malleable through biological self-testing practices within the biohacking and longevity community. Based on digital ethnographic observations of online forums and commercial age testing services, we analyse how users interpret and act upon their biological age data. Our analysis reveals that ageing is no longer viewed as an immutable biological process, but rather as a malleable combination of organic rhythms, socio-technical interventions and mundane knowledge. We refer to this as *longevity hacking*: an experimental practice of self-optimisation based on quantified ageing markers, aimed at living as long as possible, in the best possible health. This practice reconfigures notions of time, corporeality and agency, offering a new perspective on ageing as an ongoing process of maintenance and enhancement. We argue that biological age testing enables ageing to be reframed as a temporally open, controllable, and partly reversible process. This challenges the conventional view of ageing as an inevitable decline, opening up the possibility of self-directed health management grounded in molecular knowledge and collective online experimentation. We conclude that longevity hacking represents a new synthesis of ageing: a practice in which biological and socio-technological elements are interwoven into a dynamic, experimental mode of temporal and bodily modulation.

1 Introduction

Ageing and the inherent finitude of the body seem increasingly to be losing their status as biological necessities in current bioscientific discourse and the associated communities. Such debates and practices are not solely confined to tech entrepreneurs and billionaires; they have also taken root in local communities. In Germany, for instance, a variety of associations, groups and medical institutions, as well as general practitioners' surgeries, have been established alongside the Party for Rejuvenation Research, which considers longevity to be a political priority. While the individual groups and organisations differ in intensity and practice, they all draw on bioscientific knowledge and the successes

of biomedical laboratories in the field of rejuvenation to legitimise their aims. In this way, so-called ‘biohackers’ and ‘longevitists’ have been increasingly discussing medical interventions in physical ageing processes and everyday measures to slow down or even reverse biological decline at conferences and on social media. Bioinformatic models of individual ageing in organisms and the underlying sub-processes based on blood or saliva samples provide a central reference point for the new quality of imagination and practical implementation of new potential actions with regard to the prevention and intervention of the ‘meta-disease ageing’ (Spindler 2014: 44, own translation). Corresponding biomedical test procedures are also widely available on the consumer market and are enjoying increasing popularity among self-taught laypeople who, drawing on scientific debates and experience shared within their communities, are designing so-called ‘longevity protocols’, which are intended to make their own ageing available in its numerous organic manifestations – a practice we refer to as *longevity hacking*.

Longevity hacking is based on various test procedures that produce alternative age values. In contrast to the established understanding of chronological age, these values form a central reference point for technologically controlling and biologically optimising the individual ageing process using case-specific longevity practices. Our thesis is that this results in a new view of *ageing as a synthesis* of organic rhythms, sociotechnical interventions and mundane knowledge. In contrast to transhumanist ideas of an information-based mind that can be separated from the organic body using cyberware and the transformation of the human self into algorithms (cf. Ohlrogge 2025; Loh 2018), the discourses and practices of biohackers and longevitists offer a biocentric and holistic perspective (Attia and Gifford 2023; Land 2024). This perspective emphasises the preservation of human life as an interplay of physical, mental, social, spiritual and contextual factors (Sovijärvi, Arina and Land 2024). From the perspective of longevity, the pursuit of a long and healthy life renders extended lifespan constitutively linked with the malleability of bodily decline. The underlying assumption is that slowing or reversing biological ageing directly prolongs life itself. In this sense, biological ageing and longevity are intertwined in the community’s practices.

The specific characteristics of such a longevity concept include individualised health interventions based on biological analyses and data-driven practices, as well as biomedical and pharmaceutical interventions and preventive measures. Longevity hacking practices combine scientific findings, technological innovations and personal optimisation strategies. Such rejuvenation practices have long been regarded in gerontological discourse as an expression of transhumanist utopias (e.g., Dumas and Turner 2015; Fishman et al. 2008; Loh 2018; Pfaller 2016; Spindler 2014; Vincent 2009). The introduction of biomedical research into everyday life, the growing popularity of home laboratory tests and new social ideas are leading to the consolidation of a new ideal of the ‘end of ageing’ (Sinclair and La Plante 2019, own translation) among the long-lived. However, this ideal always takes the biological body into account and does not attempt

to abandon it. As Ellison aptly puts it: ‘The hacked body is the perpetually youthful, functional, and ageless body. Just as with the ‘fit body’, it is a state that can never be fully achieved and yet must constantly be striven for, a continuous propelling forward’ (Ellison 2020: 43). Here, ageing appears neither linear nor biologically determined, but rather as a temporally open process of creation – always in the making, in a continuous field of tension between rejuvenation and ageing.

Based on an online ethnography of the biohacking and longevity community, in the following article we will present their understanding of longevity and the practices of longevity hacking. First, we will outline the bioscientific understanding of ageing as a malleable process that inspires such practices (2). We will then (3) briefly outline our methodological approach and (4) empirically reconstruct that longevity hacking manifests itself through (4.1) making the biological ageing and physical decline visible, which then (4.2) opens the possibility of shaping the ageing process. Finally, we will (5) summarise our findings and argue that longevity hacking is a synthesis of ageing.

2 Ageing as a Malleable Process

Ageing fundamentally represents the process of change in a person with and through time, which can be expressed as a central measure in old age – for example, between a person's birth and the present day. This calendar-based or chronological understanding of age is socially established, but is increasingly being debated in current discussions. In particular, scientific and social debates on longevity and rejuvenation are calling into question the linear, uniform and irreversible nature of ageing. This discourse is based on a fundamental distinction between chronological age and biological age[1], a shift in the approach to preventive medicine in relation to biological ageing, and a consensus in the life sciences ‘that at least some processes involved in ageing – and perhaps more than previously thought – are modifiable’ (Sholl 2021: 6). While chronological age refers to the calendar time since birth, biological age is measured using various biomarkers and describes the physiological state of the body (Moreira 2017: 71ff.). This distinction and the plasticity attributed to biological ageing, as well as the attribution of causing disease progression or even considering ageing as a disease itself, provide the impetus for new approaches to preventive medicine that counteract age-related decline through personalised treatment protocols (Blasimme 2021: 11). The focus of preventive medicine on ageing is thus shifting from people who are already aged to young people and their predisposition to age-related diseases (Lafontaine 2015: 62). From this standpoint, ‘to

[1] In sociology, a distinction is also made between physical age, which is an individual’s perception of their own age, and social age, which is the institutionalisation of age in society and its cultural representation (van Dyk, 2020: 17).

age well […] is not to age at all' (Lafontaine 2015: 75). The concept of longevity, the idea that it is possible to delay, halt or even reverse the individual biological ageing process, is based on a bioscientific understanding of the body at a molecular level (Rose 2007). As Ellison (2019: 133) points out: 'The body of the molecular gaze is a body that is open, malleable, contingent, stochastic, and flattened; it is a body, furthermore, that destabilizes the humanist ideal of the unified and closed body of Western modernity'. In this conception, the body can be reduced to individual biophysiological elements and processes, whereby physical decay becomes increasingly separable, localisable and malleable (Cozza, Ellison and Katz 2020; Ellison 2019; Lemoine 2020; Nowotny and Testa 2009). While chronological ageing is a measure based on the passage of time, biological age is understood as a construct based on various physical parameters. The concept of longevity, as understood through bioscientific knowledge, ultimately embodies the idea of extending one's lifetime beyond the limitations of the biological ageing process. This renders ageing molecularly malleable.

In this logic, bioinformatician Aubrey de Grey claims to have discovered seven cellular and molecular causes of ageing that can be treated with regenerative medicine. Each of these causes represents a distinct damage process for which reparative approaches exist or can be theoretically developed (de Grey and Rae 2010). In this context, ageing is constituted by the combination of the various elements. While Grey's promises of healing and repair in biomedical research were dismissed as wishful thinking in the early 2000s, other research programmes have emerged over the years that, albeit more sceptical about the possibilities of human rejuvenation, are investigating regenerative approaches to treating, reversing and ending ageing processes. These programmes attribute the physical ageing process – primarily based on animal experiments – to an accumulation of various cellular and molecular damages that can be treated with medical interventions and determined along biological age values (Mykytyn 2008; Sholl 2021; Vincent 2006). The so-called 'hallmarks of ageing' (López-Otín et al. 2013, 2023) have becoming widely known in popular science. These are factors that, in their interaction, determine the ageing of organisms and include, for example, genetic damage, chronic inflammation and disrupted cell communication. The corresponding therapeutic approaches to slow down or reverse the ageing process include stem cell therapies, regenerative drug delivery and dietary interventions (López-Otín et al. 2023).

Ageing processes are studied in the laboratory using so-called biological clocks. These computer-based models measure changes in genetic material, known as DNA methylation, and compare them with age, average values and other biological characteristics (Crimmins, Klopack and Kim 2024: 1031). They form the basis for tests that determine biological age and enable statements to be made about physical condition (Pinel, Green and Svendsen 2023). Direct-to-consumer test procedures available on the market take this pattern-based approach on measurable and modifiable hallmarks and apply them to different cellular processes to generate alternative values for determining

biological age. Typically, these self-tests require the most commonly needed blood or saliva sample to be taken at home and returned to the test provider, where the biological age can be determined using the biological clocks underlying the respective method.[2] Based on these findings and drawing on scientific debates, biohackers and longevitists use experimental measures to shape their biological ageing processes and maintain their physical health in the long term. To develop and sustain the physical basis for potential future interventions, longevitists engage in various longevity practices. Many life-extension methods are based on early laboratory findings in model organisms, such as fruit flies, worms and mice. For example, calorie restriction has been shown to slow cellular ageing in mice under ideal conditions (Park 2016). Rapamycin, originally developed as an immunosuppressant, is also considered promising due to its initial rejuvenating effects in mice (Sinclair and La Plante 2019: 187). Other practices under discussion include dietary supplementation, regular exercise, consistent sleep routines, and even invasive approaches such as stem cell and gene therapies or blood plasma transfusions (Ellison 2020: 42; Sinclair and La Plante 2019: 222ff.).

Based on this understanding of age, which makes it possible to measure the ageing process using various age values and to view it as something that can be shaped, practices are increasingly emerging today that actively intervene in the ageing process and aim to achieve longevity (Ellison 2019). In addition to biological age values, digital media such as self-tracking technologies are used to test the individual effects of various interventions 'in a personalized *n=1* manner' (Swan 2012: 95, emphasise in original). The self-tracking practices of people who want to live longer by slowing down their biological ageing draw on concrete practical knowledge from scientific studies on the potential of various dietary supplements, medical interventions or nutrition-related practices to modify ageing, as well as from the quantification of their own ageing processes using biological age tests. In doing so, biohackers and longevitists experiment on their own situation, identify specific problem factors, test potential measures and modify their situation (Wettmann 2025). In such a self-experimental process of a 'reflexive self-scientification' (Zillien 2020, own translation), knowledge about success factors can be generated by successively integrating factors and measures into the experimental design, thereby transforming everyday life (Zillien, Wettmann and Peper 2023). Building on this approach to digital self-tracking as an experimental practice, our exemplary analysis of longevity hacking focuses on the question of how biological age tests contribute to the promise of individually achievable longevity.

[2] Beyond biomedical tests, biological age is increasingly embedded in everyday life: fitness trackers display a 'fitness age', smart scales calculate 'metabolic age', and health insurance companies determine an bodily lifestyle age.

3 Methodology

Our findings are based on a digital ethnographic analysis (Caliandro 2018; Hine 2015) of the online practices of longevitists who use direct-to-consumer biological age testing as part of their efforts to shape their personal longevity and pursue rejuvenation. The starting point for our analysis was an examination of digital self-tracking of sleep as part of the DFG projects ‘Sleep Knowledge: On the Production of Knowledge in Sleep Laboratories and via Self-tracking’ and ‘Sleepwalking: Recalcitrant Knowledge about a Liminal State’. In the online forums we observed, we noticed an increasing focus on longevity. As a result, in preparation for a research project on longevity, we reviewed online forums on this topic as well as publicly available advertising materials and information documents from providers of biological age tests.[3] Rather than adopting a critical perspective on commercial age testing, neoliberal health responsibilities, or medicalisation, our analysis focuses on understanding an age-related ‘culture of life’ (Knorr Cetina 2005) around biological ageing and longevity interventions. The preliminary analysis presented here therefore focuses on how results of biological age testing are negotiated, interpreted and used within the biohacking and longevity community. We therefore selected an initial corpus of 14 forum posts from the subreddits r/blueprint, r/nutrition, and r/renue. These discussions are primarily characterized by user reports on the testing process and the presentation of test results, often accompanied by self-declared interventions to improve or reverse biological ageing indicators. In addition to forum discussions, our empirical material includes one detailed report of a GlycanAge test, and two reports provided by TruDiagnostic. These test reports are complemented by five publicly available informational documents, which offer insights into the scientific foundations and evaluation procedures of the respective TruDiagnostic test kits. All selected data sources were freely accessible on the Internet, discoverable through search engines, and did not require any registration or login credentials. We exported and saved the online material for our further qualitative analysis following grounded theory (Glaser and Strauss 1967). The results presented here provide initial insights into our empirical observations regarding the use of biological age testing within the self-experimental approach of longevitists and thus mark the beginning of a joint research project on the central everyday practices, actors and debates in the context of longevity.

[3] In our further research, we would like to take a closer look at the field of longevity and its associated social worlds, including medical institutions, political organisations, tech companies, and activist groups. The following analysis is therefore our first foray into this arena.

4 Findings & Analysis

In the following analysis, we reconstruct the relational positioning of biological age testing in the context of experimentally pursued longevity along two characteristics: First, home laboratory testing of biological age is characterised by making the organic rhythms visible, which is seen as the first step in shaping ageing (4.1). Second, this results in the shaping of ageing through socio-technical and nutritional interventions, which are intended to either maintain the current physical condition or restore a previous one, with the aim of achieving longevity. (4.2). As our analysis demonstrates, ageing emerges as a synthesis construct within the context of longevity hacking.

4.1 Making Age Visible

Based on the 'molecular gaze' (Ellison 2019: 133), ageing is understood as a conglomerate of individual elements. Singular ageing processes can be measured using various biomarkers and must ultimately be synthesised to form a holistic picture. However, this also means that different tests are needed to determine the biological decay process in its entirety, such as inflammatory age, telomere age, functional age, etc. For biohackers and longevitists, the quest to shape their own biological ageing process often revolves around the measurement of biological markers and processes to determine their biological age. In the longevity community, measuring biological age is often emphasised as the starting point for individual interventions, with the aim of collecting as much data as possible. As one user puts it in an online forum: 'I decided to do the first step which was to test my biological age'. To make ageing visible, there are various commercial providers, such as TruDiagnostic. In contradistinction to preceding providers, TruDiagnostic employs myriad biological clocks to make the individual senescence process visible and thus pledges to cover the relevant organic levels of the ageing process using a solitary blood sample.[4] In addition to individual health markers, the TruDiagnostic test report provides three biological age measurements at different levels: holistic biological age (OMICmAge), organ-specific age (SYMPHONYAge) and the speed of ageing (DunedinPace). These are based on the analysis of numerous biomarkers to determine biological ageing. TruDiagnostic's home lab test thus enables the visible representation of the decay process using various biological values and scores, providing knowledge about one's own ageing, which initially leads to fragmentation of the body and the current state of decay and is then brought together through the synthesis of these ageing values. This frames ageing as a temporal biological

[4] Consequently, there is no standardised measure for the biological ageing process. Instead, measurements are subject to provider-specific definitions and methods, which leads to incompatibilities and controversial debates about validity.

effect that can be tracked along measurable current states of multiple biomarkers, rather than a calendar-based determination of age.

The visibility of ageing as a multifactorial process, as shown in the example of a chronologically 36-year-old user, can in many cases lead to confusion regarding the validity of the numerical data produced, the general state of health or the current living conditions, provided that there are significant deviations from the respective age values or with regard to the expected test results. The age values shown here for the user in question indicate consistently positive values for organ-specific age, i.e. values below or equal to chronological age. However, the OMICmAge determined by TruDiagnostic, which is considered a 'deeper reflection of your biological age', is about 10 years above his chronological age. The user comments on his results:

> Do you know why my biological age is so high ? and if it's possible to reverse it dramatically ? I've 3 hypothesis why it's so high : 1/ The TruDiagnostic test is bullshit (?) […] 2/ I might have long Covid-19 which affects the results […] 3/ I've quite a stressful Job since 10 years so could be accumulation of stress.

The user's irritations and doubts are directed at the test procedure itself, which is usually met in the community with recommendations to perform other test procedures for comparison, his own medical history and long-term effects from a previous COVID-19 infection, although his 'lung and brain age' seem to contradict this, and stress in his everyday working situation. The problematisation of individual age values is then used in discussions with the community to develop interpretation patterns from which concrete options for action are derived. Another user interprets the results as follows:

> The way that I understand it is that OMCm age reflects the lifelong experience whereas DunedinPace reflects your current habits. You're aging at a rate of 0.65 currently which is great but it's very likely that you were previously aging much more rapidly than that […] If I were you I'd focus on what you can control in there and now which is your current rate of aging.

Biological age is consequently the basis for interventions aimed at enhancing longevity, as visible signs of ageing prompt the desire to manage one's own ageing process to prolong life (Pinel, Green and Svendsen 2023). The molecular view and the biologisation of the ageing process justify an extension of the potential framework of temporal agency beyond physical decline. The reciprocal relationship between life and time thus creates a range of possibility for actively shaping ageing. In addition to the test provider's understanding of ageing as conveyed through information materials and test reports, this is evident in the interpretation of the test results and the subsequent experimental longevity practices of the longevitists, as presented and discussed in online forums and blog posts, for example. In their efforts to slow down or even prevent biological ageing, the longevitists we examined develop informational templates based on scientific debates and derive concrete practices for shaping their ageing trajectories from them.

4.2 Shapability

Making biological age visible through home lab tests leads to a practice of shaping the ageing process with the aim of longevity. The tests not only enable the current biological ageing to be recorded but also allow the success or failure of longevity hacking to be determined, i.e. measures to rejuvenate or at least maintain the current physical condition. Such measures for shaping ageing, which are intended to lead to the longest possible and healthiest life, range from dietary interventions, the use of regenerative drugs, stem cell therapies and blood transfusions. In the online forum, longevity enthusiasts discuss these options and other interventions, as well as their biological age test results, in order to design and evaluate suitable measures. This is fundamentally a self-experimental mode of testing and evaluation (Wettmann 2025) based on biological age values. Or, as one longevitist writes in the online forum: 'try and post results of biomarkers and other tests if effective – continue, if not – modify'. This sense of self-experimental testing are recurring topics in the online debates. In one thread, for example, a user presents his age tests and writes:

> Just got my TruDiagnostic results back. First test is .66 pace of aging! I'm stocked that my protocol is translating to paper. If I can get my next two test rules to be around the same pace, looks like I'll be on the leaderboard of rejuvenation Olympics. I think it goes to show that you don't need millions of dollars to achieve good results.

For the user, it becomes clear that the ageing pace in self-experimentation is considered a key indicator for visualising the individual ageing process and evaluating its slowing down as a success. While a pace of ageing of 1 indicates average ageing by one calendar year, the test result shows a slowing down of biological ageing ('.66 pace of ageing') and thus the success of age management interventions. He emphasises that, unlike publicly effective longevity millionaires such as Bryan Johnson, it does not take millions or biotechnological interventions to achieve good results. His longevity hacking is mainly based on a consistently implemented lifestyle. Factors such as sleep, nutrition and exercise are central to this. He uses the self-tracking technology Whoop to measure his sleep in order to maintain a consistent sleep duration of 8 to 8.5 hours; he eats a Mediterranean diet low in carbohydrates and high in protein; and his training includes 5 to 6 sessions per week of strength training followed by cardio training, with at least 150 minutes of Zone 2 endurance exercise, weekly runs and at least 2 minutes of Zone 5 exercise. Compared to his wife, who also wants to slow down her ageing process, he can determine the success of his protocol by determining his ageing pace:

> I'm still looking to improve where I can. This isn't a perfect protocol but right now it works for me. My wife does the same regimen, but her score came in at .94 pace of aging. She went through three months of high stress, low sleep, and bad diet before she took her test which in my opinion, gave her that score. She was taking

> all her supplements through that period. That to me means that supplements alone are not enough if you want to increase your longevity.

Even though he does not consider his protocol to be perfect, he can say that it works for him. The comparison with his wife, who has the same lifestyle but a higher ageing pace of 0.94, underlines for him that dietary supplements alone are not enough to slow down the ageing process. Despite regularly taking supplements during a period of high stress, poor sleep and poor nutrition, a negative impact on her biological age remained measurable. For him, this clearly shows that lifestyle-related and social stress factors are central to the ageing process. However, a similar finding was made by the following user:

> I've been following Bryan Johnson's supplement regimen while making some key lifestyle changes [...] The results? Pretty exciting! Telomere Length: My biological age dropped from 80.37 years in January to 63.70 years in August. DunedinPACE (Pace of Aging): Improved from 1.14 (accelerated aging) to 1.02 (close to a normal aging rate). I'm now planning to ramp up my exercise routine to see if I can push these improvements even further. The journey isn't over, but I'm excited to see where it leads!

In nine months, the user was able to slow down his ageing process by consistently changing his lifestyle, using Bryan Johnson's protocol. Specifically, his biological age, measured by telomere length, improved from 80.37 to 63.70 years; his DunedinPACE score dropped from 1.14, indicating slightly accelerated ageing, to 1.02, which is about normal. Encouraged by this success, he now wants to continue shaping his ageing process and add exercise to his protocol. He uses the measures, which are checked against measurements, to shape his everyday life in order to slow down the ageing process. In doing so, he tries to continuously improve his protocol. However, improvements here do not mean a constant upward trend and gaining complete control and availability over the ageing process. On the contrary, as the following case shows:

> Here are my results: Pace of aging = 0,68 Telemere length= 22,7 y/o. My chronological age = 24,75. Actually I slightly changed my life style. What I do the most: consume 3 table spoons of olive oil per day Collagène Glycine Curcumine I run once a week 5km I do gym once a week I have a good skincare but I dont think it has a big impact But I avoid totally sugar, eat meat once a week at most and avoid processed food (excepted bread and pasta/cheese) [...] It gives me want to continue and improve myself. [...] It seems the 80/20 rule is true.

This user was also able to improve his biological age values through minimal adjustments to his daily routine: with a chronological age of 24.75, he has a telomere age of 22.7 and an ageing pace of 0.68. He refers to the '80/20 rule', according to which 20% of targeted measures can achieve 80% of the results. Instead of radical changes, he emphasises the effectiveness of simple, consistently implemented changes such as a balanced diet, moderate exercise and avoiding unhealthy habits.

In the experimental knowledge production, protocols are constantly tested and evaluated, supplemented by successful practices of other longevitists or study-based interventions, and fundamentally changed if necessary. In this self-experimental practice of longevity hacking, an experimental self-empowerment is established, which enables users to validate predictions and promises of a longer life on their own bodies and through individually tailored practices. Longevitists share and discuss their protocols and test results in online forums, thus embody the medical promise of ageing as a malleable process, and the potential to shift previously accepted biological boundaries.

5 Discussion: Ageing as Synthesis

This analysis has revealed that biological testing methods offer new insights into ageing. The visibility and malleability of ageing are two central dimensions: first, biological tests make it possible to determine age as a biological measure and to identify, compare and correlate various physical ageing processes; second, these temporal markers open up a space of possibilities for targeted measures to slow down or even reverse the ageing process. In this dual movement, ageing is not only measured but also shaped – an interaction we refer to as longevity hacking. This refers to a practice in which ageing appears as a negotiable variable. Ageing is not viewed as exclusively biological or purely social, but rather as a relational synthesis of both: made visible through technological processes, influenced by socio-technical measures and framed by knowledge about individual, family and societal ageing. Although biological measurements serve as a yardstick, they are always relative to social knowledge about ageing and understanding of chronological age. Biological ageing can only be defined in relation to biological averages and the chronological logic of ageing. At the same time, however, chronological ageing continues to be used to interpret biological ageing. This means that concepts of age and values are interrelated. Drawing on Elias's (1992: 62) notion of 'seeing together' [*In-Beziehung-Setzen*], we understand this synthesis as a dynamic relational process in which biological, technological, and social dimensions of ageing are continuously brought into relation with one another. Ageing thus emerges as a relational composition shaped through socio-technical infrastructures, biological processes, and cultural imaginaries. In this sense, ageing no longer appears as a linear and biologically determined process, but rather as a temporally open process of shaping – always in the making, in a continuous field of tension between rejuvenation and ageing. The chronological understanding of time as a continuous and uniform flow is thus supplemented by a biological conception of time in which one's own ageing body appears as a mere temporal framework – one that contains multiple levels and layers, interacting and counteracting across diverse organic rhythms and sometimes fundamentally distinct temporalities. However, the synthetic reconstruction of biological ageing in the context of biomedical testing is by no means detached from its chronological counterpart.

6 Conclusion: The (In)Finite Nature of the Body

The biomedical deconstruction of the ageing process into multiple sub-processes, alongside the experimental appropriation of the body's own decay, gives rise to a new perspective on the finitude of the individual body. The practices of body modification and biomedical age testing, which are aimed at maintaining or even reducing biological age, are based on and further promote a shift from the inevitability of ageing to the infinity of life. The special temporal characteristic of this desired amortality is based on an intertwining of prediction and promise (Farman 2020: 29), in which the prediction of a potential slowing down or even reversal of the ageing process carries the promise of the malleability of biomedically measurable age values and is supported by this. This intertwining fits into a 'somatic sociality' (Niewöhner 2011: 291), in which biomedical knowledge about the malleability and multidimensionality of the ageing process understands and reproduces social life as a 'synthesis of nature and culture' (Wettmann and Peper 2023: 222, own translation) along its epigenetic and molecular effects on the individual body. This means that the body in its organic constitution always appears as something that has become molecular and is thus modulated by past influences, but in its molecular becoming it harbours the potential and thus the call for longevity interventions.

In the understanding of biohackers and longevitists as well as the knowledge transfer of scientific findings and biological test procedures, a view of ageing emerges as 'a contingency of evolution and not an ontological necessity' (Farman 2020: 9). This picture presupposes and enables an understanding of ageing as a combination of interdependent, yet partly antagonistic, processes that can be technologically recorded and experimentally treated. From this perspective, *longevity hacking represents a synthesis of ageing*, combining organic rhythms with technical processes, subjective embodiment, and mundane knowledge to produce an understanding of the body as malleable and temporarily open. Ageing is not abolished, but rather reimagined as a malleable project – ultimately with the aim of challenging what has been unavailable until now: 'Death is now our only foe' (Johnson 2023).

References

Attia, Peter; Gifford, Bill (2023): Outlive: The Science & Art of Longevity. New York: Harmony.

Johnson, B. (2023) Post on X (formerly Twitter), 10 November. Available at: https://x.com/bryan_johnson/status/1723028542458384540?lang=de (Accessed: 8 June 2025).

Blasimme, Alessandro (2021): The plasticity of ageing and the rediscovery of ground-state prevention. History and Philosophy of the Life Sciences 43, pp. 1–18. DOI: 10.1007/s40656-021-00414-6.

Caliandro, Alessandro. 2018. Digital Methods for Ethnography: Analytical Concepts for Ethnographers Exploring Social Media Environments. Journal of Contemporary Ethnography 47 (5), pp. 551–578. DOI: 10.1177/0891241617702960.

Cozza, Michela; Ellison, Kirsten L.; Katz, Stephen (2022): Hacking age. In Sociology Compass 16 (10). DOI: 10.1111/soc4.13034.

Crimmins, Eileen M.; Klopack, Eric T.; Kim Jung Ki (2024): Generations of epigenetic clocks and their links to socioeconomic status in the Health and Retirement Study. In Epigenomics 16 (14), pp. 1031–1042. DOI: 10.1080/17501911.2024.2373682

de Grey, Aubrey; Rae, Michael (2010): Ending Aging: The Rejuvenation Breakthroughs That Could Reverse Human Aging in Our Lifetime. New York: St. Martin's Press.

Dumas, Alex; Turner, Bryan S. (2015): Introduction: Human Longevity, Utopia, and Solidarity. The Sociological Quarterly 56 (1), pp. 1–17. DOI: 10.1111/tsq.12081.

Elias Norbert (1992): An Essay on Time. Dublin: University of Dublin Press.

Ellison, Kirsten L. (2019): Molecular Imaginaries of Aging and Age Intervention: A Discursive Analysis of Popular Science and Technology Coverage of Developments in the Field of Anti-Aging Science, Medicine, and Technology. Calgary, AB: University of Calgary.

Ellison, Kirsten L. (2020): Upgraded to Obsolescence: Age Intervention in the Era of Biohacking. In Engaging Science, Technology and Society 6, pp. 39–44. DOI: 10.17351/ests2020.361.

Farman, Abou (2020): On not dying: secular immortality in the age of technoscience. Minneapolis: University of Minnesota Press.

Fishman, Jennifer R.; Binstock, Robert H.; Lambrix, Marcie A. (2008): Anti-aging science: The emergence, maintenance, and enhancement of a discipline. Journal of Aging Studies 22 (4), pp. 295-303. DOI: 10.1016/j.jaging.2008.05.010.

Glaser, Barney; Strauss, Anselm (1967): The Discovery of Grounded Theory: Strategies for Qualitative Research. Mill Valley, CA: Sociology Press.

Hine, Christine. 2015. Ethnography for the Internet: Embedded, Embodied and Everyday. London; New York: Bloomsbury Academic. DOI: 10.4324/9781003085348.

Lafontaine, Céline (2010): Die postmortale Gesellschaft. Wiesbaden: VS Verlag für Sozialwissenschaften.

Land, Siim (2024): The Longevity Leap: A Guide to Slowing Down Biological Aging and Adding Healthy Years to Your Life. Independently published.

Lemoine, Maël (2020): Defining Aging. In Biology & Philosophy 35 (46), pp. 1–30. DOI: 10.1007/s10539-020-09765-z.

Loh, Janina (2018): Trans- und Posthumanismus zur Einführung. Hamburg: Junius Verlag.

López-Otín, Carlos; Blasco, Maria A.; Partridge, Linda; Serrano, Manuel; Kroemer, Guido (2013): The Hallmarks of Aging. In Cell 153, pp. 1194–1217. DOI: 10.1016/j.cell.2013.05.039.

López-Otín, Carlos; Blasco, Maria A.; Partridge, Linda; Serrano, Manuel; Kroemer, Guido (2023): Hallmarks of aging: An expanding universe. In Cell 186, pp. 243–278. DOI: 10.1016/j.cell.2022.11.001.

Moreira, Tiago (2017): Science, Technology and the Ageing Society. London: Routledge.

Mykytyn, Courtney E. (2008): Medicalizing the optimal: Anti-aging medicine and the quandary of intervention. Journal of Aging Studies 22 (4), pp. 313–32. DOI: 10.1016/j.jaging.2008.05.004.

Niewöhner, Jörg (2011): Epigenetics: Embedded bodies and the molecularisation of biography and milieu. BioSocieties 6 (3), pp. 279–298.

Nowotny, Helga; Testa, Giuseppe (2009): Die gläsernen Gene: Die Erfindung des Individuums im molekularen Zeitalter. Frankfurt: Suhrkamp.

Ohlrogge, Carsten (2025): Nach dem Menschen? Die Frage nach dem Tod im Transhumanismus und die Grenzen sozialer Bezugnahme. In: Benkel, Thorsten; Meitzler, Matthias (eds.): Jahrbuch für Tod und Gesellschaft 4, pp. 118–133. Weinheim: Beltz Juventa.

Park, Hyung W. (2016): Old Age, New Science: Gerontologists and Their Biosocial Visions. Pittsburgh: University of Pittsburgh Press.

Pfaller, Larissa (2016): Anti-Aging als Form der Lebensführung. Wiesbaden: Springer VS.

Pinel, Clémence; Green, Sara; Svendsen, Mette N. (2023): Slowing down decay: biological clocks in personalized medicine. Frontiers in Sociology 8, pp. 1–13. DOI: 10.3389/fsoc.2023.1111071.

Rose, Nikolas (2007): The Politics of Life Itself: Biomedicine, Power, and Subjectivity in the Twenty-First Century. Princeton, NJ: Princeton University Press.

Sholl, Jonathan (2021): Can aging research generate a theory of health?. History and Philosophy of the Life Sciences 43 (45), pp. 1–26. DOI: 10.1007/s40656-021-00402-w.

Sinclair, David; LaPlante, Matthew D. (2019): Lifespan: Why We Age—and Why We Don’t Have To. New York: Atria Books.

Sovijärvi, Olli; Arina, Teemu; Land, Siim (2024): The Resilient Being: Mastering the Biology of Stress & Resilience. Tallinn: Hololife Publishing.

Spindler, Mone (2014): ‘Altern ja – aber gesundes Altern’: Die Neubegründung der Anti-Aging-Medizin in Deutschland. Wiesbaden: Springer VS.

Swan, Melanie (2012): Health 2050: The Realization of Personalized Medicine through Crowdsourcing, the Quantified Self, and the Participatory Biocitizen. Journal of Personalized Medicine 2, (3) pp. 93–118. DOI: 10.3390/jpm2030093.

van Dyk, Silke (2020): Soziologie des Alters. 2., aktualisierte und ergänzte Ausgabe. Bielefeld: transcript.

Vincent, John A. (2006): Ageing Contested: Anti-ageing Science and the Cultural Construction of Old Age. Sociology 40 (4), pp. 681–698. DOI: 10.1177/0038038506065154.

Vincent, John A. (2009): Ageing, Anti-ageing, and Anti-anti-ageing: Who are the Progressives in the Debate on the Future of Human Biological Ageing? Medicine Studies 1 (3), pp. 197–208. DOI: 10.1007/s12376-009-0016-6.

Wettmann, Nico. 2025, in press. Sleep Tracking. Zur Genese des ‘guten’ Schlafs. Weinheim: Beltz Juventa.

Wettmann, Nico; Peper, Frederik (2023): Schlaf als Synthese. Soziotechnische und temporale (Re-)Figurationen digitaler Schlafvermessung. In: Benkel, Thorsten; Meitzler, Matthias (eds.): Mythenjagd Soziologie mit Norbert Elias, pp. 220–240. Weilerswist-Metternich: Velbrück Wissenschaft.

Zillien, Nicole (2020): Digitaler Alltag als Experiment Empirie und Epistemologie der reflexiven Selbstverwissenschaftlichung. Bielefeld: transcript Verlag.

Zillien, Nicole; Wettmann, Nico; Peper, Frederik (2023): Sleep Experiments. Knowledge Production through Self-Tracking. Historical Social Research 48 (2), pp. 157–175. DOI: 10.12759/HSR.48.2023.20.

Co-creating Systemic Knowledge about Community Acceptance: Guidance for integrating Causal Loop Diagrams and Participatory System Mapping in Acceptance Research

Marius Rogall[1], Jan-Hendrik Kamlage[1 2], David Sasse[1], Klaus Krumme[2 3]

[1]Ruhr-University Bochum, Germany

[2] University of Duisburg-Essen, Germany

[3] National Technical University/ Kharkiv Polytechnic Institute, Institute of Education and Science in Economics, Management and International Business, Ukraine

DOI 10.3217/978-3-99161-062-5-006, CC BY 4.0

Abstract. The energy transition is a key component in achieving Germany's and Europe's environmental and energy policy goals. While public support for the energy transition is generally high, local conflicts surrounding the related infrastructure projects are slowing down the transformation and causing costs to rise. Acceptance research focuses on the factors and contextual conditions under which such projects are accepted or rejected by affected stakeholders. However, previous research approaches are lacking systemic perspectives that consider the interactions of factors in locally specific constellations. In this article, we develop a conceptual framework that enables us to analyse complex local constellations of acceptance formation. Our approach combines systemic and participatory perspectives on community acceptance of renewable energy technologies (RET) and translates them into a systematic methodological approach in the form of causal loop diagrams (CLD) and participatory system mapping (PSM). The potential of this methodology is illustrated using preliminary results from a case study on electricity grid expansion. These show that CLDs are suitable for capturing, visualising and understanding complex causal mechanisms in the process of acceptance formation. Due to the collaborative research process of researchers and stakeholders within the PSM, the results show an increased relevance for the implementation of communication strategies in the local context. Overall, the combination of systemic and participatory research methods in the form of CLDs and PSM is a suitable approach to expand the methodology and analytical framework of acceptance research. It enables complexity to be captured and thus advances our understanding of acceptance formation.

1. Introduction

The energy transition is a key component in achieving Germany's and Europe's environmental and energy policy goals (Bundesregierung, 2023; Europäische Kommission, 2019). The expansion of renewable energies and their integration into the existing energy system represents the central challenge in this regard (Bertsch et al., 2016). This process of transformation manifests itself tangibly in the form of energy infrastructures, such as wind turbines, electricity pylons, large-scale transformers and ground-mounted PV systems (Kühne, 2024; Walker, 2024; Weber, 2019).

The support of the German population for the expansion of these renewable energy technologies (RET) and for the energy transition in general has been consistently high for many years (Bertsch et al., 2016; Setton, 2020). However, in communities affected by the construction of energy infrastructure, conflicts and resistance arise frequently, as the burdens of change become visible and the landscape is transformed (Devine-Wright and Devine-Wright, 2009). Local protests and resistance consistently lead to increased costs and delays in the realisation of RET projects (Löschel et al., 2013).

The acceptance and non-acceptance of infrastructure projects forms at the level of people's individual motives or attitudes and exists on a continuum between the two poles of approval and rejection of a project. On this continuum, positions vary between active support, simple approval or tolerance to complete rejection.[1] Acceptance is fragile and the result ‘of a complex, permanent process of communication and action between acceptance subjects and acceptance objects extending over the entire life cycle of an acceptance object’ (Bentele et al., 2015, p. 5). In addition to the political and social factors that influence acceptance, research is increasingly focussing on structural and spatial conditions, such as the value of landscape or place attachment (Delcayre and Bourdin, 2025; Devine-Wright, 2009).

However, there is a lack of research that (1) adopts a systemic perspective on the complex and dynamic local acceptance formation processes, (2) translates this into a systematic methodology and (3) links it with participatory research approaches in order to validate the findings discursively against the practice of local stakeholders.

Our contribution addresses this research gap. We propose the combination of two complementary methods. Through the integrated use of Causal Loop Diagrams (CLD) and Participatory System Mapping (PSM), we develop a holistic and systemic methodological approach that takes into account the complexity and context-sensitive

[1] The term acceptance is often insufficiently defined in research on energy infrastructures and renewable energies and is often barely differentiated from similar terms such as support, resistance, uncertainty or apathy (Batel et al., 2013).

formation of community acceptance, incorporates discursively validated practical knowledge and thus generates socially robust findings (Nowotny et al., 2001).

In the following chapter, we present our conceptual framework for analysing the acceptance of RET in affected communities with the help of CLDs and PSM. We have translated this conceptual framework into a concrete methodological approach as part of a case study from the German electricity grid expansion. We will explain this approach in more detail in Chapter 3. In Chapter 4, we illustrate the possible results of our approach on the basis of examples from the case study. Finally, Chapter 5 reflects on the gains and challenges of the proposed research procedure.

2. Conceptual Framework

In their framework, Wüstenhagen et al. (2007) distinguish between three central and interwoven dimensions of acceptance: socio-political, community and market acceptance. While the dimension of socio-political acceptance addresses general support in politics and society, market acceptance refers primarily to economic and market players. In this article, we focus on the third dimension of community acceptance. This refers to acceptance of various stakeholder groups at the local level, such as residents, local entrepreneurs, local politicians, and local clubs and initiatives. The level of community acceptance depends on the attitudes of these local actors with regard to a new technology or infrastructure that is realised in the immediate proximity. It is the result of a complex interplay of diverse factors from the local and superordinate spatial levels (Wolsink, 2018). Kluskens et al. elaborate on this idea by understanding local acceptance formation as a process of weighing up different objects of acceptance. Stakeholders at the community level evaluate, for example, the location of an infrastructure or the planning process and finally arrive at an overall assessment of the project. That means, even in cases where there is no active resistance to the project, not all of these aspects are necessarily accepted, i.e. ‘even in the unproblematic cases acceptance is ambiguous’ (Kluskens et al., 2024, p. 842).

In previous research, there is a knowledge gap with regard to such consideration processes and the interaction between various influencing factors. Previous studies have mainly focused on identifying individual factors relating to specific problems (for an overview, see Kamlage et al., 2024), like landscape changes and their effects on place attachment and place identity of the affected community (Devine-Wright and Devine-Wright, 2009; Kühne, 2018), psychological issues like risk/benefit evaluations, trust and perceived fairness (Gross, 2007; Huijts et al., 2012; Richter et al., 2016), public information and participation (Kamlage et al., 2020) or the role of community benefits and financial participation (Cowell et al., 2011; Schönauer and Glanz, 2023).

In contrast, the 'fertile ground' approach by Delcayre und Bourdin (2025) offers a more valuable analytical approach to address the complexity of the local acceptance formation process. In their view, community acceptance largely depends on the extent to which the project characteristics are compatible with a series of ‘territorial characteristics’. These are defined as various specific local factors like socio-economic structures, place attachment, past experiences and historical lines of conflict (Delcayre and Bourdin, 2025).

However, in general, research to date has mostly lacked a methodology that integrates systemic perspectives on local acceptance formation and can thus capture complexity and context instead of reducing or ignoring them. In terms of methodology the predominantly used research methods are quantitative surveys (Huijts et al., 2007; Zoellner et al., 2008) or qualitative methods such as expert interviews, media analyses and participant observation in the context of case studies (Sanchez Nieminen and Laitinen, 2025). While qualitative case studies can capture the complex constellations at least descriptively, through a dense and inductive description of specific cases (e.g. Eichenauer and Gailing, 2022; Fienitz, 2025), it is difficult for studies with a quantitative survey method (Baxter et al., 2013; Hoen et al., 2019; Zoellner et al., 2008) to overcome the isolated consideration of individual factors.

In order to address the lack of systemic perspectives and to comprehend this process of weighing up different factors, we have developed a conceptual framework that translates a systemic perspective on community acceptance into a systematic methodological approach and also integrates participatory research methods (see Figure 1). To capture, visualise and understand the various acceptance factors, their relationships and the complex and dynamic interaction patterns that emerge, we use CLDs as a methodological tool of the system thinking approach (Forrester, 1968; Sterman, 2004). When developing the CLDs, we use the methodology of Participatory System Mapping (PSM) which integrates relevant stakeholders into the research process (Barbrook-Johnson and Penn, 2022). The mutual validation in the dialogue between researchers and stakeholders minimises subjective bias in the construction of the CLDs and increases the epistemic quality of the results. This procedure offers a twofold gain in knowledge: Systemic depth and local, contextualised relevance. In the following, we describe the methodology of CLDs and PSM in more detail.

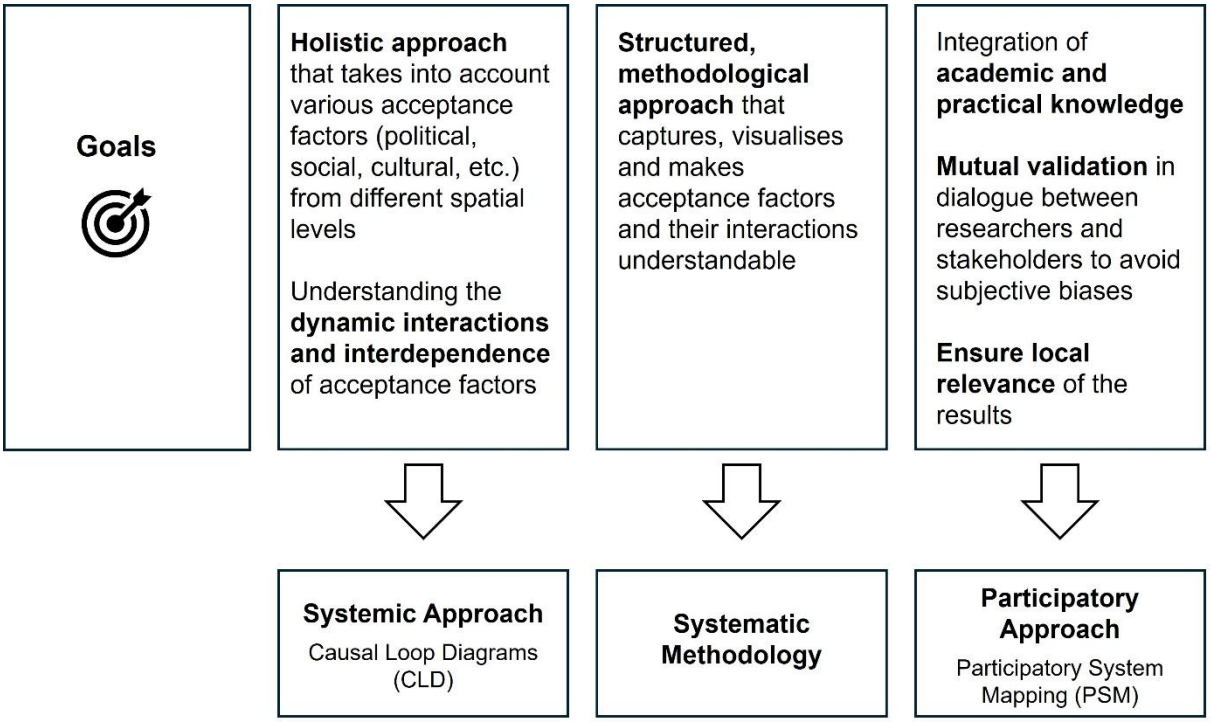


Fig. 1: Illustration of our conceptual framework, authors' own presentation.

2.1. Causal Loop Diagrams (CLD)

CLDs are a tool for visualising causal relationships between different elements of a system. CLDs consist of three core components. Firstly, the variables - in our case the acceptance factors. These acceptance factors are interwoven through causal links, which are the second core component. These links have a polarity, which is indicated by + or - . A + means that both variables change in the same direction. A - on the other hand indicates that both variables are moving in opposite directions. This is illustrated by the examples in Figure 2. If the number of citizen initiatives (CIs) increases, public attention for the power line project also increases, or if the number of CIs decreases, public attention for the power line project also decreases (same direction). If the number of existing infrastructure increases, the amount of available land decreases, or if the number of existing infrastructure decreases, the available land increases (opposite direction).

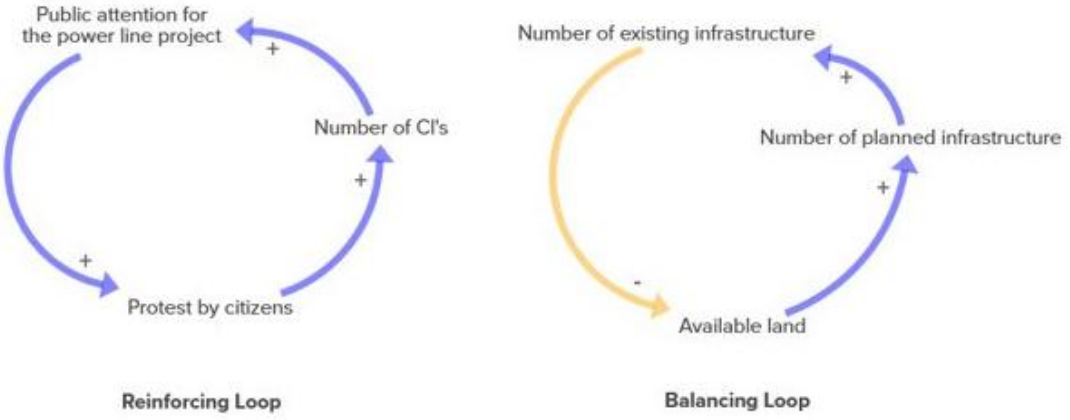


Fig. 2: Examples for a reinforcing and a balancing loop, authors' own presentation. Legend: + / violet arrows = same direction; - / orange arrows = opposite direction

The third component of a CLD are the feedback loops. A distinction is made between reinforcing loops, which represent an exponential development, and balancing loops, which represent an equalising development or an approach to a state of equilibrium. In the example given in figure 1 for a reinforcing loop, the increasing number of CIs leads to an increase in public attention for the project, which reinforces the protest, which in turn increases the number of CIs. The variables reinforce each other and protest mobilisation increases, as the rising attention for the project involves more actors who organise themselves into CIs and thus generate more attention. The dynamics surrounding infrastructure-related land use conflicts, on the other hand, are an example for a balancing loop. A high availability of land in a municipality creates incentives to plan and realise infrastructural projects. The higher the number of planned infrastructure the higher the number of actually existing infrastructure. However, the more infrastructure exists, the less land is available and with little land available few new infrastructure projects are going to be planned.

With the help of CLDs, we are able to take a systemic perspective on our research topic of community acceptance. Factors from different social subsystems (political, social, cultural, etc.) and spatial levels (local, regional, national, global) can be integrated into the CLD and related to each other. The visual form of the CLDs also makes it possible to consider a large number of factors simultaneously without reducing the complexity of the case (Barbrook-Johnson and Penn, 2022). This allows feedback loops and other system-dynamic mechanisms to become visible.

2.2. Participatory System Mapping (PSM)

The complex systems of acceptance formation can be assessed comprehensively and plausibly through the collaboration of researchers and stakeholders. Kates et al. very early pointed out that: 'participatory procedures involving scientists, stakeholders, advocates, active citizens, and users of knowledge are critically needed' (Kates et al.,

2001, p. 641). Lang et al. argue that complex real world phenomena and problems need the constructive knowledge inputs of various affected societal groups and perspectives to be relevant for the practice (Lang et al., 2012, p. 25f.). According to Norström and others, co-production processes should be context based and locally embedded, pluralistic and inclusive, goal-oriented and interactive in nature (Norström et al., 2020).

To properly represent this productive and collaborative basic understanding, we integrated participatory research methods into our conceptual framework. We used the Participatory System Mapping (PSM) method (Barbrook-Johnson and Penn, 2022), to develop a CLD through a participatory process with stakeholder, reflecting local acceptance formation in our case study.

First coined as a formal method by Sedlacko et al. (2014) in the context of knowledge brokerage on sustainable consumption, PSM has since diversified rapidly and has been used in several sustainability related domains, such as last-mile logistics and local food networks (De La Torre et al., 2019; Gruchmann et al., 2019; Melkonyan et al., 2017), tourism policy design (Suno Wu et al., 2021; Tourais and Videira, 2021), ecosystem-service governance in marine coastal zones (Lopes and Videira, 2017, 2015), business sustainability in rural dairy enterprises (Kamath et al., 2019), and transport-decarbonisation strategies (Penn et al., 2022). Collectively, these applications demonstrate how the original CLD-based workshop format has become a versatile, stakeholder-centred tool for tackling complex sustainability challenges across multiple domains.

In our collaborative research methodology, we draw on deliberative design principles (Niemeyer et al., 2024) to conceptualise a process that facilitates transparent, open and free discourse on acceptance factors, while minimising the effects of interpersonal power structures. Such a process enables the discursive validation of problem structures and system understandings. The perspectives and validity claims that come up during the process are based on shared and mutually recognised arguments. From this perspective intersubjectively confirmed knowledge is not discovered, but co-produced under conditions that promote communicative rationality (Habermas, 1981; Thompson, 1983).

3. Implementation in Methodology

Based on the conceptual framework described above, we have developed a concrete methodological approach in the context of a case study from the German electricity grid expansion in order to investigate community acceptance in relation to the construction of a new power line.

There are templates in the literature for structuring PSM workshops with stakeholders and generally for the iterative process of PSM from collaborative mapping workshops and post-production phases of the researchers (Barbrook-Johnson and Penn, 2021,

2022; Lopes and Videira, 2015; Sedlacko et al., 2014). Usually, the first step involves a joint workshop of researchers and stakeholders to jointly develop an initial draft of the CLD. For pragmatic considerations and against the background of experience from a previous case study, we decided to deviate from this proposal. Instead of starting the first workshop with a blank sheet, we created a first draft of the CLD as part of the case study described here on the basis of qualitative data collected by us and validated and further developed this in discourse with stakeholders in the PSM workshop. This decision was primarily made due to time constraints on the part of the stakeholders involved - the representatives of the Transmission System Operator (TSO), which is responsible for the planning, construction and subsequent operation of the power line. We had 3.5 hours available for the PSM workshop. As the development and discussion of a CLD is very time-consuming and methodologically demanding due to the complex interrelationships in the social systems under consideration, there is a risk that a workshop for the joint construction of an initial version of the CLD will fail due to excessive demands on the stakeholders involved and will end with results that are of little use and biased.

Accordingly, our research process is divided into the following four phases (see Figure 3): (1) Drafting a CLD; (2) Conducting a PSM workshop; (3) Iterative feedback and further development of the CLD (editing); (4) Final Workshop. The individual phases are explained in more detail below.

3.1. Drafting a CLD

Prior to the first draft of the CLD, a comprehensive process of data collection, evaluation and analysis took place. The first step involved collecting a large amount of qualitative data (see Table 1). This came from participant observations at TSO information and participation events in affected municipalities. In addition, various text documents were analysed, including articles from the local press, statements from local stakeholders and websites of protest actors.

Source		Number
Documents	TSO Statements	9
	Public Media / Press	93
	Political Publications	7
	Social Media Posts	7
	Formal Statements	4
	CI websites	12
Observations	Information events of the TSO	6

Table 1: Overview of the empirically collected qualitative data in the case study presented here.

The resulting empirical material was analysed and coded with regard to the identification of acceptance factors and relationships between these factors. In a second step, these were transferred to a cross table.

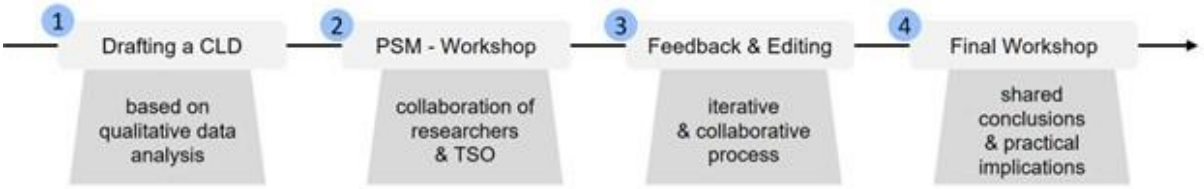


Fig. 3: Illustration of the research process, authors' own presentation.

A total of 49 different variables and 68 relationships between them were identified. A first version of the CLD was developed on the basis of the cross table, which served as the basis for the PSM.

This included a total of six central mechanisms. We define mechanisms here as a construct of relationships between various interdependent local factors which, in their specific combination, have an effect on the acceptance of the power line project. One of these central mechanisms is the core engine of our CLD. The core engine is the centrepiece of a CLD. It forms the basis from which the entire diagram is developed and expanded (Barbrook-Johnson and Penn, 2022). In our case, the core engine is the mechanism that represents the process of protest mobilisation against the planned power line in the form of a reinforcing loop (see Chapter 4).

3.2. PSM-Workshop

The next step was to organise a PSM workshop, which plays a central role in our research process. In addition to the researchers, six representatives of the TSO took part. The workshop served to validate the first draft of the CLD from the perspective of the stakeholders involved in a transparent, open, inclusive and moderated collaboration process and to develop it further in the discourse. The perspectives, assumptions and validity claims of the researchers and TSO representatives were critically reflected upon and mutually acknowledged in dialogue in order to arrive at a common understanding of the acceptance formation process under consideration (Lopes and Videira, 2015). The dialogue about the CLD deepened the understanding of the acceptance formation process among all participants and opened up new perspectives.

An introduction to the CLD methodology is essential in order to enable stakeholders to participate constructively in the workshop. For this reason, the TSO representatives were introduced to the methodology and systemic perspectives on acceptance in advance of the workshop, and the workshop itself also began with a brief introduction to the perspectives of system thinking and the syntax of CLD. In this way, a basic understanding of the method and thus a basis for discussion for the content part was created.

In order to avoid overwhelming the stakeholders in the workshop with the extensive CLD, it was sent to them in advance along with some introductory information to aid understanding. In addition, the six central mechanisms of the CLD were explained step by step by the researchers during the workshop, and previous steps and assumptions in the research process were made transparent. The mechanisms were then discussed separately with the stakeholders and gradually linked together. This enabled the stakeholders to develop a good understanding of the CLD. During the workshop, the mechanisms and the CLD were projected onto the wall using a projector and also laid out on the table in printed form. The comments, questions and additions that arose during the workshop were recorded in written form on the printed copy of the CLD. Finally, the workshop participants were given the opportunity to prioritise certain variables, relationships or sub-areas of the CLD.

3.3. Iterative feedback and further development of the CLD (editing)

Following the PSM workshop, the comments and additions collected there were processed and incorporated into the CLD (post-production phase). In some cases, this also meant more in-depth research, the results of which were incorporated into the CLD in the form of new variables and improved the analytical depth. In total, 5 new factors and 11 new links were incorporated into the CLD following the workshop. One variable from the first draft of the map was removed. The revised CLD will be returned to the workshop participants from the TSO with a request for further feedback. Depending on the amount of comments and questions, this can be done by email or in an online meeting.

The CLD will be further developed in the course of the case study. This will be done in close cooperation with the representatives of the TSO in the form of recurring feedback meetings and subsequent post-production phases in which the comments will be incorporated.

3.4. Final Workshop

The case study concludes with another workshop with the TSO representatives. Here, the final version of the CLD is discussed and validated once again. The aim of this workshop is to bring together the results of the participatory process and the qualitative data collected in a final, plausible and coherent CLD, whose conclusions are shared by all participants. Furthermore, effective points of intervention are to be identified in a joint discussion that can have a decisive influence on local acceptance. These intervention points may then be incorporated into the TSO's future communication strategy and addressed where possible.

The workshop will conclude a process lasting several months, during which researchers and representatives of the TSO developed shared knowledge about local acceptance formation in a specific case study and recorded it in the form of a CLD and implications for practice.

4. Empirical Implications

The following section illustrates the results of the approach described in the previous chapter and the insights that CLDs can provide. For this purpose, simplified excerpts from the CLD developed in our case study are shown below as examples. The case study has not yet been finalised, which is why the following illustrations and conclusions do not claim to be complete or conclusive.

Figure 4 shows the central mechanism, or core-engine, of our CLD.

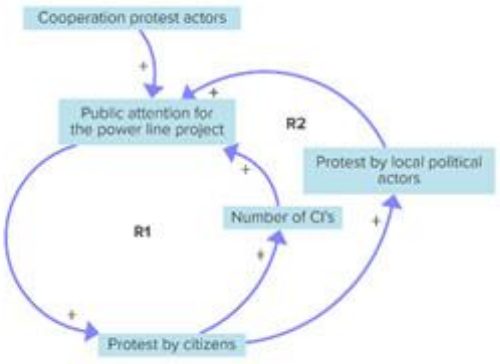


Fig. 4: Loop of protest mobilisation (core engine) from the CLD of the case study presented here, authors' own presentation. Legend: + / violet arrows = same direction; blue marked factors = factors of the core engine

It consists of a reinforcing loop that depicts the dynamics of protest mobilisation in our case study. Due to the difficulties in empirically capturing the often tacit acceptance of infrastructure projects such as power lines, we worked with a negative definition and used the CLD to map how constellations of factors affect the dynamic development of protest against the project.

The reinforcing loop R1 describes the mechanism by which citizens join the protest, organise themselves in the form of CIs and thus draw the attention of a wider public to the issue and their position. This attracts new members to the protest movement and so on. This feedback loop is reinforced by the fact that other actors, such as the affected municipalities, join the protest alliance and thus give it further attention and legitimacy. This can lead to an exponentially growing protest mobilisation. For us, the research question linked to this dynamic is: What factors reinforce or hamper this reinforcing loop of protest mobilisation?

In order to answer our research question, we first identified various primary factors and linked them to the variables of our core engine (see Figure 5). By primary factors, we mean those factors that have a direct influence on the factor ‘protest by citizens’ and thus on the protest mobilisation loop. ‘Local burdens’ associated with the new power line, ‘doubts about the need’ of the power line and the perceived ‘threat to local identity’ intensify the protest and thus also drive the protest mobilisation loop. Perceived procedural fairness, on the other hand, tends to lead to greater acceptance of the project and can slow down the loop of protest mobilisation.

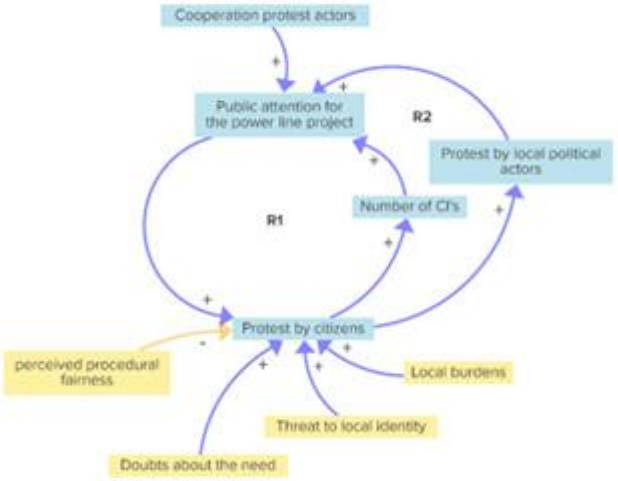


Fig. 5: Loop of protest mobilisation (core engine) and primary factors from the CLD of the case study presented here, authors’ own presentation. Legend: + / violet arrows = same direction; - / orange arrows = opposite direction; blue marked factors = factors of the core engine; yellow marked factors = primary factors

However, in order to go beyond the identification of direct influencing factors, further secondary factors were identified and discussed, which are linked to each other and to the primary factors and thus have an indirect effect on the loop of protest mobilisation (see Figure 6).

This clarifies further modes of action that explain the relevance of the primary factors identified in this case study and were part of the driving force that has driven protest mobilisation on a large scale in our case study so far.

It is also worth taking a closer look at these secondary factors in order to identify possible leverage points in the system of local acceptance formation. These are ‘places to intervene in a system’ (Meadows, 1999). In other words, these are the acceptance factors in our system that have a particularly strong influence on the rest of the system and on protest mobilisation. Candidates for these intervention points can be found among the factors that have a high out-degree, i.e. that themselves influence many other factors, but at the same time are themselves only influenced by a few other factors - i.e. have a low in-degree (Kiekens et al., 2022).

In the simplified representation of our CLD in Figure 6, the factors ‘bundling with other infrastructure’ and ‘traceability of planning decisions’ are important candidates for effective leverage points (marked in green). Both factors are not influenced by any other factors in the system shown, but themselves influence other key factors.

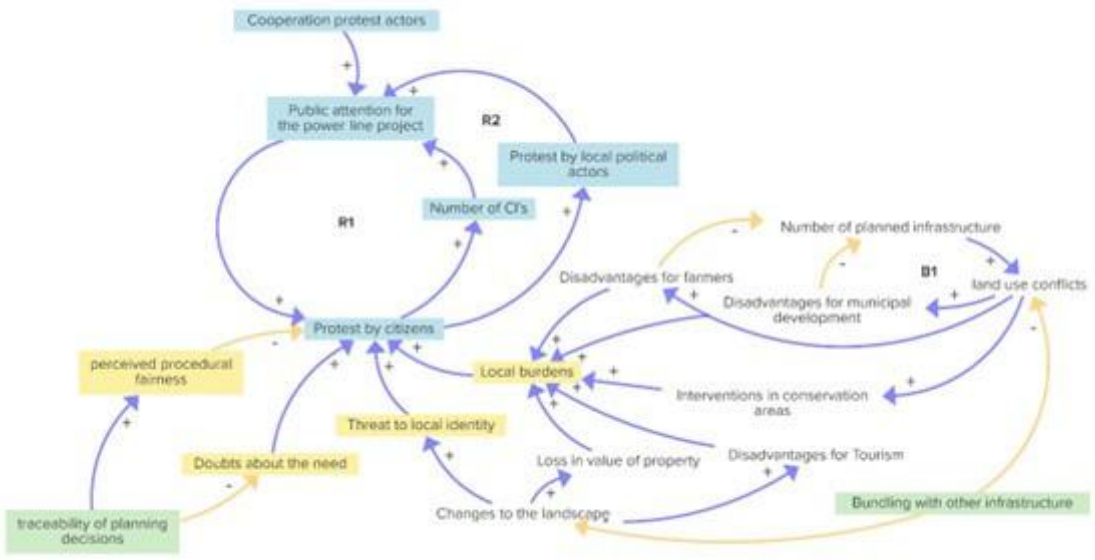


Fig. 6: Simplified and condensed version of the CLD from the case study presented here, authors’ own presentation. Legend: + / violet arrows = same direction; - / orange arrows = opposite direction; blue marked factors = factors of the core engine; yellow marked factors = primary factors; green marked factors = potential leverage points

The ‘traceability of planning decisions’ has an impact on two primary factors (‘perceived procedural justice’ and ‘doubts about the need’), which have a direct influence on our

core engine - the protest mobilisation loop. The ‘bundling with other infrastructure’, on the other hand, has an influence on the ‘changes to the landscape’ and the ‘land use conflicts’. These are both factors that are themselves potential points of intervention, as they each influence three other factors in the system.

Based on the (simplified) CLD illustrated here, it can be hypothesised that actors such as the TSO could mitigate the protest against the planned power line or increase acceptance of the project by striving to justify all planning decisions in a comprehensible manner and planning the line as often as possible in close proximity to other existing or planned infrastructure. These hypotheses can be discussed well in the context of PSM workshops with the stakeholders involved and developed into concrete intervention measures (Barbrook-Johnson and Penn, 2021; Sedlacko et al., 2014).

In any case, the importance of some secondary factors for the genesis of local acceptance formation processes is likely to be overlooked or at least underestimated without systematic analysis with the help of CLDs.

5. Discussion

Our Goal was to enrich research on community acceptance by expanding it with systemic and participatory methods and perspectives and to integrate them into a systematic methodology. In this final chapter, we reflect on the extent to which the methodological approach proposed here can fulfil this goal and what challenges and limitations need to be considered.

Previous research has mainly identified bundles of individual factors influencing acceptance, without paying sufficient attention to the dynamic interaction patterns between the various factors that cause escalating protests and conflicts. Overall, we believe that CLDs are indeed a suitable method to capture, visualize and understand these complex interaction patterns between a large number of acceptance factors. They therefore offer a valuable analytical addition to previous acceptance research. The inductive approach of our methodology allows us to integrate factors from different social subsystems (political, social, cultural, economic, etc.) and spatial levels (local, regional, national, global) into the analysis and to relate them to each other. It does not set any thematic boundaries. From the analysis of various text documents and the discussions in PSM workshops, a large number of acceptance factors and links were identified. The visual representation of the CLD makes it possible to include them all in the analysis at the same time. It is not uncommon for CLDs to contain between 20 to 50 or even more different variables (Barbrook-Johnson and Penn, 2022). CLDs make it possible to visualise system-dynamic modes of action which have a decisive influence on the acceptance formation process, such as feedback loops. Identifying these dynamic mechanisms is important in the context of analysing community acceptance, as it allows

us to better understand, for example, rapid escalations of protest and conflict. Furthermore, CLDs offer the potential to identify effective intervention points for influencing the system of community acceptance. The potential impact of these ‘leverage points’ can only be recognised from a systemic perspective because they are ‘often not intuitive’ (Meadows, 2008, p. 147).

The attempt to analyse acceptance from a systemic perspective with the help of CLDs is not entirely new. Ketzer et al. (2020) used a system dynamics approach to analyse factors that affect the acceptance of agro-photovoltaic systems and González et al. deal with the acceptance of renewable energy projects in poor rural communities (González et al., 2016). However, these studies rely exclusively on scientific literature or their own empirical data when constructing their CLDs, which poses the risk that the subjective perspective of the researchers causes distortions in the CLDs and its analysis (Barbrook-Johnson and Penn, 2022). To avoid these distortions and to increase the validity of our CLD we worked with a participatory system mapping (PSM) approach. The central component of this is the collaborative work of researchers and representatives of the TSO on the CLD in a workshop setting where an open discourse based on mutual recognition takes place and a common understanding of the local acceptance formation process is developed.

At the end of this participatory process stands a CLD based on judgements that were collaboratively and discursively validated by researchers and TSO representatives. The findings on acceptance formation that are produced by this process are highly relevant for the local context of the analysed case of a municipality affected by the expansion of the electricity grid.

Our experience with the approach outlined above also shows that researchers need to consider and reflect on a number of challenges. These are both methodological and theoretical in nature. The methodological challenges include (1) involving all relevant stakeholder groups in PSM and (2) the time-consuming iterative nature of the participatory process.

For the best possible result, it is recommended to integrate several or all stakeholder groups into the PSM (Barbrook-Johnson and Penn, 2022). Nevertheless, our case shows that this is not always possible, especially if stakeholders are on different and possibly opposed sides of a latent or even manifest conflict. These potentially conflicting framework conditions must be reflected in the research design and its implementation. The fact that we only involved representatives of the TSO is a significant limitation of our approach that leads to distortions in the representation of the system of acceptance formation. At least some perspectives of local stakeholders were gained through participant observation at information events and document research. These were included in the draft of the CLD, but the contents of the CLD are still only validated from two perspectives (TSO and researcher). However, it was not possible to speak with the

TSO representatives in a protected atmosphere in any other way. In workshops with other stakeholders, they would have been less open about their perspectives and assessments. That is why we decided to focus solely on the workshops with the TSO representatives in order to first test and further develop the methodology of CLDs and PSM for acceptance research. With a more validated methodology, further PSM workshops with other stakeholder groups should also be conducted in the future to obtain a comprehensive picture of local acceptance formation that includes various relevant perspectives.

Our previous experience has also taught us that developing CLDs within the framework of PSM takes up a lot of time, both for the researchers and the stakeholders involved. During the workshops, sufficient time is needed for discussion and the development of a common understanding. In addition, stakeholders must be carefully introduced to system thinking and the methodology. Outside of the workshops, familiarisation with the topic and the iterative process of revision and feedback also require a significant amount of time. The literature on PSM estimates the time required for a PSM workshop at between 80 minutes (Sedlacko et al., 2014), 3 hours (Barbrook-Johnson and Penn, 2021) and 4 hours (Lopes and Videira, 2015). In our experience, however, this is not enough time to adequately introduce the stakeholders to the CLD methodology and to collaboratively develop an initial draft of the CLD. This was the main reason for creating an initial draft of the CLD based on our empirical data, before involving the stakeholders in the design process. With this pragmatic decision, certain limitations are created. The idea behind PSM is that CLDs are 'intended to be 'owned' by the stakeholders who create them, rather than researches' (Barbrook-Johnson and Penn, 2022, p. 64). By creating the first draft of the CLD as a research team at the forefront of the workshop, we shifted ownership of the CLD away from the stakeholders and channelled the discussion in a certain direction.

The theoretical challenges of CLDs include (1) the temporary validity of evidence and (2) their lack of generalisability. Although CLDs are sometimes characterised as mere 'snapshots' of a system at a single point in time (Sedlacko et al., 2014, p. 36), they actually embed temporal information implicitly through feedback loops, delays and accumulations. Classic system-dynamics archetypes such as Limits to Growth or Shifting the Burden (all expressed solely as CLDs) capture characteristic time-dependent behaviours including exponential growth, overshoot-and-collapse, and path dependence (Senge, 1990; Sterman, 2004). What CLDs cannot provide on their own is a quantitative trace of when those behaviours will manifest; translating the diagram into a stock-and-flow model or complementing it with longitudinal evidence (e.g., process tracing) is necessary to generate testable predictions (Sterman, 2004). Moreover, because variable selection and boundary assumptions are context-specific, the explanatory power of any given CLD remains tied to the socio-ecological conditions under which it was constructed (Sedlacko et al., 2014). Future research should therefore pair participatory CLDs with

explicitly process-oriented methods (such as repeated mapping sessions, sequence analysis or process tracing) - to examine how stable the depicted feedback structure remains as contextual factors evolve (see for an empirical example Fienitz, 2025).

Because the causal-loop diagrams (CLDs) generated by our participatory procedure encode the perceptions of a particular community, the insights they yield are inherently context-specific and not directly generalisable (Sedlacko et al., 2014). At the same time, every CLD uses the same syntactic elements - variables, signed causal links, and feedback loops - so maps from different cases can be systematically compared as long as their boundaries and variable names are documented consistently (Lane and Oliva, 1998). Comparative work of this kind has already uncovered a set of recurring feedback configurations known as system archetypes that appear across very different domains (Kim, 2000; Senge, 1990). As more case studies of local acceptance for energy-infrastructure projects are visualised as CLDs, future research could search for such archetypal patterns to identify feedback structures that repeatedly shape community responses to energy infrastructure. Doing so would strengthen the external validity of individual maps and provide theory-informed leverage points for stakeholder engagement.

6. Conclusion

Despite extensive scholarship on community acceptance, researchers still lack a convincing explanation of how interacting social, institutional and spatial factors and feedbacks determine whether RET-projects are welcomed or resisted. By adopting a participatory systems lens, our study addresses this gap.

Using Participatory Systems Mapping in a case study from the electricity grid expansion, we engaged representatives of the regional TSO to co-develop a causal-loop diagram (CLD) that makes the interdependencies among various factors such as trust, perceived fairness, landscape attachment and procedural efficacy explicit. This collaborative process enabled mutual validation of perspectives from practice and science and the development of a common and in-depth understanding of the local acceptance formation process.

Although the absence of municipal stakeholders inevitably biases the current CLD towards the operator’s viewpoint, the exercise demonstrates that systemic, participatory modelling can enrich acceptance research by capturing complexity rather than reducing it. Future iterations should iterate the CLD with local residents, local politicans and NGOs, enabling cross-case comparison and the identification of recurring system archetypes that shape acceptance dynamics across projects.

References

Barbrook-Johnson, P., Penn, A., 2021. Participatory systems mapping for complex energy policy evaluation. Evaluation 27, 57–79. https://doi.org/10.1177/1356389020976153

Barbrook-Johnson, P., Penn, A.S., 2022. Systems Mapping. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-031-01919-7

Batel, S., 2020. Research on the social acceptance of renewable energy technologies. Energy Research & Social Science 68, 101544. https://doi.org/10.1016/j.erss.2020.101544

Batel, S., Devine-Wright, P., Tangeland, T., 2013. Social acceptance of low carbon energy and associated infrastructures: A critical discussion. Energy Policy 58, 1–5. https://doi.org/10.1016/j.enpol.2013.03.018

Baxter, J., Morzaria, R., Hirsch, R., 2013. A case-control study of support/opposition to wind turbines: Perceptions of health risk, economic benefits, and community conflict. Energy Policy 61, 931–943. https://doi.org/10.1016/j.enpol.2013.06.050

Bentele, G., Bohse, R., Hitschfeld, U., Krebber, F., 2015. Akzeptanz in der Medien- und Protestgesellschaft – Gedanken, Analysen, Thesen, in: Bentele, G., Bohse, R., Hitschfeld, U., Krebber, F. (Eds.), Akzeptanz in der Medien- und Protestgesellschaft. Springer Fachmedien Wiesbaden, Wiesbaden, pp. 1–22. https://doi.org/10.1007/978-3-658-06167-8_1

Bertsch, V., Hall, M., Weinhardt, C., Fichtner, W., 2016. Public acceptance and preferences related to renewable energy and grid expansion policy: Empirical insights for Germany. Energy 114, 465–477. https://doi.org/10.1016/j.energy.2016.08.022

Bundesregierung, 2023. Gesetz für den Ausbau erneuerbarer Energien (Erneuerbare-Energien-Gesetz - EEG 2023).

Cowell, R., Bristow, G., Munday, M., 2011. Acceptance, acceptability and environmental justice: the role of community benefits in wind energy development. Journal of Environmental Planning and Management 54, 539–557. https://doi.org/10.1080/09640568.2010.521047

De La Torre, G., Gruchmann, T., Kamath, V., Melkonyan, A., Krumme, K., 2019. A System Dynamics-Based Simulation Model to Analyze Consumers' Behavior Based on Participatory Systems Mapping – A 'Last Mile' Perspective, in: Melkonyan, A., Krumme, K. (Eds.), Innovative Logistics Services and Sustainable Lifestyles. Springer International Publishing, Cham, pp. 165–194. https://doi.org/10.1007/978-3-319-98467-4_8

Delcayre, H., Bourdin, S., 2025. In Search of ‘Fertile Ground’: How Territorial Characteristics Influence the Social Acceptability of Renewable Energy Projects. Environmental Management 75, 867–882. https://doi.org/10.1007/s00267-025-02113-5

Devine-Wright, H., Devine-Wright, P., 2009. Social representations of electricity network technologies: Exploring processes of anchoring and objectification through the use of visual research methods. British J Social Psychol 48, 357–373. https://doi.org/10.1348/014466608X349504

Devine-Wright, P., 2009. Rethinking NIMBYism: The role of place attachment and place identity in explaining place-protective action. Community & Applied Soc Psy 19, 426–441. https://doi.org/10.1002/casp.1004

Eichenauer, E., Gailing, L., 2022. What Triggers Protest? - Understanding Local Conflict Dynamics in Renewable Energy Development. Land 11. https://doi.org/10.3390/land11101700

Europäische Kommission, 2019. Der europäische Grüne Deal.

Fienitz, M., 2025. How do land use conflicts escalate? Identifying causal mechanisms in a conflict over a biogas plant in Brandenburg, Germany. People and Nature pan3.70038. https://doi.org/10.1002/pan3.70038

Forrester, J.W., 1968. Principles of systems text and workbook chapters 1 through 10, 2. preliminary ed. ed. Wright-Allen, Cambridge, Mass.

González, A., Sandoval, H., Acosta, P., Henao, F., 2016. On the Acceptance and Sustainability of Renewable Energy Projects—A Systems Thinking Perspective. Sustainability 8, 1171. https://doi.org/10.3390/su8111171

Gross, C., 2007. Community perspectives of wind energy in Australia: The application of a justice and community fairness framework to increase social acceptance. Energy Policy 35, 2727–2736. https://doi.org/10.1016/j.enpol.2006.12.013

Gruchmann, T., De La Torre, G., Krumme, K., 2019. Mapping Logistics Services in Sustainable Production and Consumption Systems: What Are the Necessary Dynamic Capabilities?, in: De Boer, L., Houman Andersen, P. (Eds.), Operations Management and Sustainability. Springer International Publishing, Cham, pp. 223–246. https://doi.org/10.1007/978-3-319-93212-5_12

Habermas, J., 1981. Theorie des kommunikativen Handelns. Suhrkamp, Frankfurt am Main.

Hoen, B., Firestone, J., Rand, J., Elliot, D., Hübner, G., Pohl, J., Wiser, R., Lantz, E., Haac, T.R., Kaliski, K., 2019. Attitudes of U.S. Wind Turbine Neighbors: Analysis of a Nationwide Survey. Energy Policy 134, 110981. https://doi.org/10.1016/j.enpol.2019.110981

Huijts, N.M.A., Midden, C.J.H., Meijnders, A.L., 2007. Social acceptance of carbon dioxide storage. Energy Policy 35, 2780–2789. https://doi.org/10.1016/j.enpol.2006.12.007

Huijts, N.M.A., Molin, E.J.E., Steg, L., 2012. Psychological factors influencing sustainable energy technology acceptance: A review-based comprehensive framework. Renewable and Sustainable Energy Reviews 16, 525–531. https://doi.org/10.1016/j.rser.2011.08.018

Kamath, V., Biju, S., Kamath, G.B., 2019. A Participatory Systems Mapping (PSM) based approach towards analysis of business sustainability of rural Indian milk dairies. Cogent Economics & Finance 7, 1622172. https://doi.org/10.1080/23322039.2019.1622172

Kamlage, J., Uhlig, J., Rogall, M., Warode, J., 2024. Shaping Energy Landscapes: Public Participation and Conflict Resolution in Wind Power, Grid Expansion, and Biogas Transformation Fields, in: Berr, K., Koegst, L., Kühne, O. (Eds.), Landscape Conflicts. Springer Fachmedien Wiesbaden, Wiesbaden, pp. 281–310.

Kamlage, J.-H., Drewing, E., Reinermann, J.L., De Vries, N., Flores, M., 2020. Fighting fruitfully? Participation and conflict in the context of electricity grid extension in Germany. Utilities Policy 64, 101022. https://doi.org/10.1016/j.jup.2020.101022

Kates, R.W., Clark, W.C., Corell, R., Hall, J.M., Jaeger, C.C., Lowe, I., McCarthy, J.J., Schellnhuber, H.J., Bolin, B., Dickson, N.M., Faucheux, S., Gallopin, G.C., Grübler, A., Huntley, B., Jäger, J., Jodha, N.S., Kasperson, R.E., Mabogunje, A., Matson, P., Mooney, H., Moore, B., O'Riordan, T., Svedin, U., 2001. Sustainability Science. Science 292, 641–642. https://doi.org/10.1126/science.1059386

Ketzer, D., Schlyter, P., Weinberger, N., Rösch, C., 2020. Driving and restraining forces for the implementation of the Agrophotovoltaics system technology – A system dynamics analysis. Journal of Environmental Management 270, 110864. https://doi.org/10.1016/j.jenvman.2020.110864

Kiekens, A., Dierckx de Casterlé, B., Vandamme, A.-M., 2022. Qualitative systems mapping for complex public health problems: A practical guide. PloS one 17, e0264463. https://doi.org/10.1371/journal.pone.0264463

Kim, D.H., 2000. System Archetypes I, Toolbox reprint series. Pegasus Communications, Cambridge, Mass.

Kluskens, N., Alkemade, F., Höffken, J., 2024. Beyond a checklist for acceptance: understanding the dynamic process of community acceptance. Sustain Sci 19, 831–846. https://doi.org/10.1007/s11625-024-01468-8

Kühne, O., 2024. Landscape and Conflict—Some Basic Considerations, in: Berr, K., Koegst, L., Kühne, O. (Eds.), Landscape Conflicts, RaumFragen: Stadt – Region – Landschaft. Springer Fachmedien Wiesbaden, Wiesbaden, pp. 19–40. https://doi.org/10.1007/978-3-658-43352-9_2

Kühne, O., 2018. ‚Neue Landschaftskonflikte' – Überlegungen zu den physischen Manifestationen der Energiewende auf der Grundlage der Konflikttheorie Ralf Dahrendorfs, in: Kühne, O., Weber, F. (Eds.), Bausteine Der Energiewende. Springer Fachmedien Wiesbaden, Wiesbaden, pp. 163–186. https://doi.org/10.1007/978-3-658-19509-0_8

Lane, D.C., Oliva, R., 1998. The greater whole: Towards a synthesis of system dynamics and soft systems methodology. European Journal of Operational Research 107, 214–235. https://doi.org/10.1016/S0377-2217(97)00205-1

Lang, D.J., Wiek, A., Bergmann, M., Stauffacher, M., Martens, P., Moll, P., Swilling, M., Thomas, C.J., 2012. Transdisciplinary research in sustainability science: practice, principles, and challenges. Sustain Sci 7, 25–43. https://doi.org/10.1007/s11625-011-0149-x

Lopes, R., Videira, N., 2017. Modelling feedback processes underpinning management of ecosystem services: The role of participatory systems mapping. Ecosystem Services 28, 28–42. https://doi.org/10.1016/j.ecoser.2017.09.012

Lopes, R., Videira, N., 2015. Conceptualizing Stakeholders' Perceptions of Ecosystem Services: A Participatory Systems Mapping Approach. Environmental and Climate Technologies 16, 36–53. https://doi.org/10.1515/rtuect-2015-0011

Löschel, A., Flues, F., Pothen, F., Massier, P., 2013. Der deutsche Strommarkt im Umbruch: Zur Notwendigkeit einer Marktordnung aus einem Guss. Wirtschaftsdienst 93, 778–784. https://doi.org/10.1007/s10273-013-1598-x

Meadows, D.H., 2008. Thinking in systems. Chelsea Green Publ, Vermont.

Meadows, D.H., 1999. Leverage points-places to intervene in a system. The Sustainability Institute,.

Melkonyan, A., Krumme, K., Gruchmann, T., De La Torre, G., 2017. Sustainability assessment and climate change resilience in food production and supply. Energy Procedia 123, 131–138. https://doi.org/10.1016/j.egypro.2017.07.236

Niemeyer, S., Veri, F., Dryzek, J.S., Bächtiger, A., 2024. How Deliberation Happens: Enabling Deliberative Reason. Am Polit Sci Rev 118, 345–362. https://doi.org/10.1017/S0003055423000023

Norström, A.V., Cvitanovic, C., Löf, M.F., West, S., Wyborn, C., Balvanera, P., Bednarek, A.T., Bennett, E.M., Biggs, R., De Bremond, A., Campbell, B.M., Canadell, J.G., Carpenter, S.R., Folke, C., Fulton, E.A., Gaffney, O., Gelcich, S., Jouffray, J.-B., Leach, M., Le Tissier, M., Martín-López, B., Louder, E., Loutre, M.-F., Meadow, A.M., Nagendra, H., Payne, D., Peterson, G.D., Reyers, B., Scholes, R., Speranza, C.I., Spierenburg, M., Stafford-Smith, M., Tengö, M., Van Der Hel, S., Van Putten, I., Österblom, H., 2020. Principles for knowledge co-production in sustainability research. Nat Sustain 3, 182–190. https://doi.org/10.1038/s41893-019-0448-2

Nowotny, H., Scott, P., Gibbons, M., 2001. Re-thinking science: Knowledge and the public in an age of uncertainty, Online-Ausg. ed. Polity, Cambridge, England Malden, Mass.

Penn, A.S., Bartington, S.E., Moller, S.J., Hamilton, I., Levine, J.G., Hatcher, K., Gilbert, N., 2022. Adopting a Whole Systems Approach to Transport Decarbonisation, Air Quality and Health: An Online Participatory Systems Mapping Case Study in the UK. Atmosphere 13, 492. https://doi.org/10.3390/atmos13030492

Richter, I., Danelzik, M., Molinengo, G., Nanz, P., Rost, D., 2016. Bürgerbeteiligung in der Energiewende. IASS Working Paper.

Sanchez Nieminen, G., Laitinen, E., 2025. Understanding local opposition to renewable energy projects in the Nordic countries: A systematic literature review. Energy Research & Social Science 122, 103995. https://doi.org/10.1016/j.erss.2025.103995

Schönauer, A.-L., Glanz, S., 2023. Local conflicts and citizen participation in the German energy transition: Quantitative findings on the relationship between conflict and participation. Energy Research & Social Science 105, 103267. https://doi.org/10.1016/j.erss.2023.103267

Sedlacko, M., Martinuzzi, A., Røpke, I., Videira, N., Antunes, P., 2014. Participatory systems mapping for sustainable consumption: Discussion of a method promoting systemic insights. Ecological Economics 106, 33–43. https://doi.org/10.1016/j.ecolecon.2014.07.002

Senge, P.M., 1990. The fifth discipline: the art and practice of the learning organization, 1. ed. ed, A currency book. Doubleday Currency, New York.

Setton, D., 2020. Soziale Nachhaltigkeit Wagen – Die Energiewende aus Sicht der Bevölkerung: Eine umfassende Auswertung der Daten des Sozialen Nachhaltigkeitsbarometers der Energiewende 2017 und 2018 mit den Schwerpunkten gerechte Kostenverteilung, Windausbau an Land sowie Digitalisierung und Verbraucherpräferenzen. Institute for Advanced Sustainability Studies (IASS). https://doi.org/10.2312/IASS.2020.007

Sterman, J.D., 2004. Business dynamics systems thinking and modeling for a complex world, Internat.ed. ed, Simulation software and models including ithink, Powersim, and Vensim software. McGraw-Hill, Boston [u.a.

Suno Wu, J., Barbrook-Johnson, P., Font, X., 2021. Participatory complexity in tourism policy: Understanding sustainability programmes with participatory systems mapping. Annals of Tourism Research 90, 103269. https://doi.org/10.1016/j.annals.2021.103269

Thompson, J.B., 1983. Rationality and Social Rationalization: An Assessment of Habermas's Theory of Communicative Action. Sociology 17, 278–294. https://doi.org/10.1177/0038038583017002010

Tourais, P., Videira, N., 2021. A participatory systems mapping approach for sustainability transitions: Insights from an experience in the tourism sector in Portugal. Environmental Innovation and Societal Transitions 38, 153–168. https://doi.org/10.1016/j.eist.2021.01.002

Walker, B., 2024. Energy-Landscape Conflicts and the Politics of Scale Around Photovoltaic Parks in Germany, in: Berr, K., Koegst, L., Kühne, O. (Eds.), Landscape Conflicts, RaumFragen: Stadt – Region – Landschaft. Springer Fachmedien Wiesbaden, Wiesbaden, pp. 335–349. https://doi.org/10.1007/978-3-658-43352-9_18

Weber, F., 2019. Stromnetzausbau und Landschaft, in: Kühne, O., Weber, F., Berr, K., Jenal, C. (Eds.), Handbuch Landschaft, RaumFragen: Stadt – Region – Landschaft. Springer Fachmedien Wiesbaden, Wiesbaden, pp. 871–883. https://doi.org/10.1007/978-3-658-25746-0_70

Wolsink, M., 2018. Social acceptance revisited: gaps, questionable trends, and an auspicious perspective. Energy Research & Social Science 46, 287–295. https://doi.org/10.1016/j.erss.2018.07.034

Wüstenhagen, R., Wolsink, M., Bürer, M.J., 2007. Social acceptance of renewable energy innovation: An introduction to the concept. Energy Policy 35, 2683–2691. https://doi.org/10.1016/j.enpol.2006.12.001

Zoellner, J., Schweizer-Ries, P., Wemheuer, C., 2008. Public acceptance of renewable energies: Results from case studies in Germany. Energy Policy 36, 4136–4141. https://doi.org/10.1016/j.enpol.2008.06.026

Trust and Manipulation in Generative AI: A Digital Humanist Perspective*

Francesco Striano[1], Maria Zanzotto[1 2]

[1] University of Turin, Italy

[2] Northwest Italy Philosophy PhD Program – FINO, Italy

DOI 10.3217/978-3-99161-062-5-007, CC BY 4.0

Abstract. This paper explores the dynamics of trust and manipulation in generative AI systems, proposing digital humanism as a critical framework to re-evaluate our relationship with such technologies. We conceptualise trust as an evaluative act – a normative judgement about the trustworthiness of a system in a given context – and argue that trust in generative AI is structurally misguided. This is not because such systems lack moral agency, but because the trust placed in them has been uncritically extended from deterministic technologies, whereas generative models are probabilistic and non-linear. These systems should be approached not as 'truth-tellers', but as 'storytellers.' We further argue that deceptive features – such as their anthropomorphic linguistic style and confident rhetorical tone – exacerbate this misalignment, making users more vulnerable. Digital humanism offers a fruitful perspective for understanding these dynamics, encouraging us to engage with AI not as neutral tools, but as cultural artefacts that shape our values, behaviour, and epistemic practices.

1 Introduction

Trust in technology is often based on an implicit model: data are entered, a system processes them automatically, and delivers reliable results. This paradigm, which has emerged in the context of deterministic systems – where mechanisms are readable, behaviours predictable, and errors traceable – has been uncritically extended to more complex and inherently opaque technologies. Generative systems and especially large language models (LLMs) deviate significantly from this model. Their mode of operation is not aimed at verifying the truth of a statement, but at producing results that are

* The paper is the result of scientific discussion and collaboration between the authors, was conceived in a joint effort and revised together. For the purposes of identifying the parts, where required, it is specified that sections 1, 2, and 4 are to be attributed to Francesco Striano, while sections 3 and 5 to Maria Zanzotto.

statistically plausible, coherent in context, and rhetorically effective. It is about a shift from *truth production* to *story production*.

In this paper, we argue that the misplaced extension of trust to generative technologies distorts our understanding of how they work and reinforces specific epistemic and political vulnerabilities. Generative AI is capable of producing content that is often indistinguishable from that created by humans. This capability undermines users' epistemic agency – their ability to critically evaluate, contextualise, and validate information. As a result, the use of these systems can erode individual autonomy – especially when deployed in high-density communicative environments – and compromise the conditions for democratic deliberation, particularly in environments where political opinion formation is already characterised by opaque platform dynamics.

However, recognising the manipulative potential of generative AI also opens up space for critical reflection. Rather than advocating uncritical trust or categorical rejection, we argue for a situated engagement with these technologies – one that emphasises interpretive consciousness and reflexive interaction. This perspective is in line with the ethos of *digital humanism*, which sees technology not merely as a neutral tool, but as a cultural and ethical phenomenon embedded in, and formative of, human values.

Building on this framework, the paper is structured in three parts. First, we will examine the misplaced extension of trust to probabilistic technologies. In doing so, we will focus on how generative AI challenges notions of reliability, trust, and confidence by creating 'stories' rather than stating facts. Secondly, looking at the gap between what these technologies *seem to be doing* and what *they do* we will discuss deception and how it can influence the evaluative act of trust, with potential consequences for epistemic agency. Finally, we will apply the perspective of digital humanism to the promotion of digital literacy in order to encourage a more critical understanding and conscious interaction with these technologies and their outcomes.

2 From Predictability to Plausibility: A Conceptual Shift in Technological Trust

Trust in information circulating through digital platforms has long been based on a more fundamental trust in the technologies that mediate it. While digital systems have sometimes provoked scepticism or outright rejection – reminiscent of historical patterns of technophobia and resistance – the prevailing tendency, particularly in Western contexts, has been to regard them as reliable infrastructures. This perceived reliability has often served as the basis for a broader attribution of trustworthiness. To clarify what is at stake in this attribution, and to understand why this trust may no longer hold in the

context of generative AI, we begin by disentangling three key concepts: *reliability*, *trust*, and *confidence*[1].

A technologically mediated society depends on the functional autonomy of its components – whether human, mechanical, informational, or hybrid[2]. However, this autonomy is never absolute: it requires and is maintained by varying degrees of trust. As Mariarosaria Taddeo notes, 'a society in which there is no trust in doctors, teachers, or drivers' would require all individuals to invest significant resources in constant monitoring, diverting time and attention from their own tasks (Taddeo, 2017, p. 566). In this view, trust enables coordination without constant monitoring and ensures that complex systems function without falling into recursive control loops.

The question of how to define trust – especially in relation to artificial agents – has led to a broad and unsettled debate. In her work, Taddeo (2010) defines trust as a second-order property that characterises binary, goal-oriented relationships: a trustor chooses to pursue a given outcome through the capacity of a trustee who is perceived to be trustworthy. This perception transforms the relationship into one that is expected to be beneficial to the trustor. Such a model has the advantage of being easily applicable to artificial systems, especially if they are designed to fulfil delineated functions with measurable success criteria.

However, this definition is not without limitations. Firstly, it assumes a binary relationship that does not readily accommodate distributed forms of trust as we see in institutions, infrastructures, or socio-technical ecosystems. Second, the definition harbours the danger of circular reasoning: it states that trust is justified by the trustworthiness of the trustee, but the criteria by which this trustworthiness is determined remain under-defined. This ambiguity becomes particularly pressing in the case of technologies, that lack moral motivation or intentionality in the human sense.

To clarify this issue, it is useful to distinguish between trust, reliance, and confidence. Several scholars have argued that what is often referred to as 'trust' in technologies is rather a form of reliance (Blackburn, 2010; Thompson, 2018). According to de Fine Licht and Brülde (2021), reliance is a three-place relation in which an agent A relies on B to achieve an outcome C. Reliance can be voluntary or involuntary, and can be directed toward both persons and artefacts. For example, we may rely on a wristwatch to tell the right time – not because we attribute some form of moral agency to it (such as a commitment not to lie), but because we judge it to be mechanically sound and consistent with our previous experience.

---

[1] For this non-standard conception of trust and a broader discussion of it, see Striano (2024b).

[2] For a discussion on the possibility of artificial agents with proper agency, see among others, Calì (2023), Floridi (2023), Himma (2009), Striano (2024a), and Swanepoel (2021).

In contrast, trust involves a specific kind of agential reliance in which the trustor ascribes some form of normative responsibility to the trustee (de Fine Licht and Brülde, 2021). This moral dimension is what makes betrayal possible: one can be betrayed by a human, but not by a machine. As Baier (1986) notes, failed reliance results in disappointment, while betrayed trust leads to moral injury. On this basis, some authors argue that technologies cannot truly be trusted because they lack intentionality and moral interest (Deley and Dubois, 2020; Thompson, 2018). They claim that we trust the designers, developers, or institutions behind the technology. In this sense, the reliability of a device serves as a proxy for trust in its makers.

Yet this distinction, while analytically useful, does not fully capture the phenomenology of trust in technological environments – especially, as we will see, when it comes to our interactions with conversational agents, whose apparent responsiveness invites forms of trust that go beyond purely functional reliance.

As Shionoya (2001) suggests, trust can be better understood as an *evaluative act*: a judgement by one agent regarding whether another – human or not – is trustworthy under specific conditions. This view considers both interpersonal and socio-technical forms of trust and opens up the space to consider trust not just as a property, but as a dynamic practice. Shionoya also emphasises the role of confidence as a disposition to trust: a background state of openness or readiness that enables the evaluative act.

Building on this insight, we propose a tripartite scheme:

- Confidence is an underlying disposition that leads a trustor to make an evaluation act of trust;
- Trust is an evaluative act that judges on the trustworthiness of the trustee;
- Trustworthiness is the characteristic trait that the trustor considers the trustee (human or artefact) to have.

Inversely, we can describe reliance in these terms:

- Reliability is a characteristic disposition of a person or an artefact;
- Reliance is the evaluation act based on the observable reliability of a person or an artefact;
- Confidence is a disposition inspired by repeated judgments of reliance (and can, in turn, lead to trust).

This scheme allows us to explain how trust can be extended to artefacts without anthropomorphising them. An artefact that consistently exhibits reliable behaviour can become the object of an evaluative act of trust – not because it possesses a will or a moral agency, but because it is *invested* with trustworthiness by the user. In this sense, trust becomes a normative stance, not a descriptive attribution.

Ultimately, both trust and reliance are grounded in confidence, but they differ in their normative assumptions. Where reliance involves functional expectation, trust implies a moral orientation. However, while it is legitimate to invest artificial systems with a form of moral trust based on evaluative judgement, this investment becomes problematic when applied to generative models such as large language models (LLMs). The difficulty lies not in their artificial nature *per se*, but in the erroneous extension of a trust model derived from deterministic technologies to systems whose functioning is qualitatively different.

Much of our habitual reliance on digital technologies has been shaped by interaction with systems governed by linear, deterministic processes. These technologies typically implement procedures that could in principle be carried out by humans – albeit more slowly – through explicit rules, formal logic, or algorithmic deduction. In such cases, the output is the result of a controlled and interpretable transformation of the input data. In knowledge cultures strongly shaped by scientific rationalism, this has contributed to a general association between digital computation, correctness, and truth. Science and Technology Studies remind us that such association has always been mediated by delegation and black-boxing rather than direct access to underlying mechanisms (Latour, 1987; Star, 1999; Edwards, 2010). Yet in the case of digital systems, this delegation rests on a genuinely linear and reproducible architecture: input, process, and output can, at least in principle, be traced and verified. Our trust in these technologies, while socially mediated, is also supported by their reliability, i.e., their consistent performance within a rationalist paradigm of control and predictability.

However, the trust we place in LLMs often assumes that they belong to the same category of reliable and explainable systems. These models work according to different principles. While they are technically deterministic at the code and infrastructure level, their output is generated by probabilistic models trained on large data sets. Their architecture introduces contingency, reflexivity, and a certain degree of unpredictability into user interaction. They are not designed to produce verified truths or facts: while they can draw on factual data, the outputs they produce do not represent facts in a direct sense, but rather construct plausible *narratives* in response to prompts. These models are not optimised for the production of truth, but for the continuation of interaction – through responses that are syntactically coherent, semantically persuasive, and rhetorically open-ended.

While we describe LLMs as narrative producers, this should not be understood as an attribution of narrative *intentionality* or autonomous meaning-making. The narratives they produce emerge from statistical coherence rather than interpretive intent. Yet meaning can still arise in the interaction between the model's patterned coherence and the user's interpretive engagement. In this sense, LLMs do not generate meaning as such, but rather *afford* it – they offer discursive structures that invite human interpretation.

This interactive production of coherence, however, should not be mistaken for epistemic reliability or truth orientation. The meaning that emerges in human-machine exchanges remains contingent on interpretation, not verification. As such, LLMs do not simply answer, but *simulate* attitudes, positions, and modes of discourse. Their persuasive power often masks the lack of epistemic commitment. In this respect, they resemble what Harry Frankfurt famously called a *bullshitter*: a speaker who does not care whether what they say is true or false as long as it serves their purpose. Drawing on Frankfurt, Gorrieri (2024) identifies three criteria for bullshit: indifference to truth-values, lack of acknowledgement of this indifference, and an ulterior communicative goal.

All three criteria, Gorrieri argues, seem to be applicable to systems such as ChatGPT. First, the model has no internal mechanism for assessing the truth-value of its outputs: it merely predicts plausible token sequences. Second, while disclaimers such as 'ChatGPT may produce incorrect information' now appear in the user interface, they frame the issue as an error rather than a structural indifference to truth. Third, the system encourages continued engagement – it ends its outputs with follow-up questions or invitations for further elaboration – showing that it is optimised not for accuracy but for sustained interaction.

This behaviour has significant normative implications. Users often interpret syntactically fluent and rhetorically sophisticated output as epistemically reliable, a misconception that is reinforced by interface design and interaction dynamics. The problem is not that such systems deliberately lie, but that they simulate a truth-oriented discourse without any concern for truth. When design choices systematically encourage users to conflate plausibility with reliability and coherence with truth, trust is not only misplaced – it is structurally misguided. In this sense, even if artificial agents cannot morally 'betray' us, we can still speak of a betrayal of trust by design.

3 Trust as a relational concept: how deceiving human-like features of generative AI pose additional issues to trust

Beyond this structural misplacement, trust in generative AI also needs to be understood as a relational phenomenon. The way users engage with LLM-based chatbots – through natural language and human-like cues – introduces additional layers to the evaluative act of trust. These systems mimic human-like characteristics, which often leads users to anthropomorphise them, meaning an attribution of human capabilities or mental states that chatbot do not possess. It is in this gap between appearance and reality that the notion of deception arises. In this section, we will explore how these deceptively human-like features can have additional effects on trust and thus on epistemic agency, i.e., the control epistemic agents have on belief-formation and belief-revision processes (Schlosser, 2019).

When people talk about chatbots deceiving us, they are usually talking about chatbots taking over the world and destroying humanity. An example of the concern about the destruction of humanity by AI is a statement on the risk of extinction caused by AI signed by prominent figures in Silicon Valley, such as Sam Altman (CEO of OpenAI), Demis Hassabis (CEO of Google DeepMind), as well as many academics and other technology leaders including Bill Gates (Hinton et. al., 2023). This overconfident attitude towards the power of AI also emerges when it comes to the capacity of AI to deceive, which is usually treated as its ability to trick humans in the pursuit of its goals, by means that do not align with human values. This discrepancy in values raises the fear that AI would be able to resort to any means to achieve its goals, even to the destruction of humans if necessary. However, this is not the direction we want to take. In fact, it is useful to distinguish between two levels of AI: general AI (also known as AGI: Artificial General Intelligence) and strong AI vis-a-vis narrow AI and weak AI. Those who speak of AI taking over the world have in mind a vaguely specified technological entity that has the same cognitive capabilities as humans, such as the ability to form mental states (intention, sentience, etc.) – strong AI –, and an intelligence that enables it to perform any kind of task – general AI –, ultimately better than humans (superintelligence). This vision is closer to sci-fi than to computer science. What computer scientists have been able to develop so far, however, is narrow AI, i.e., specialised AI systems developed for specific tasks, and weak AI, a simulation of intelligence, rather than duplication. ChatGPT, the most widely used LLM-based chatbot, is an example of narrow and weak AI that has its basis in the discipline known as NLP (Natural Language Processing), i.e., the field of computer science that deals with the development of models and systems capable of producing or modifying human-like text or speech (in this category we find chatbots, autocorrects, translators, etc.). LLM-based chatbots are more sophisticated because their computational engines are the most powerful: they use deep learning to make connections. Of course, being a narrow or weak AI does not mean that the outputs are not good, but they are fundamentally different systems from AGI. LLMs are extremely powerful linguistic machines that, as said before, calculate the most probable sequence of words given the input prompt. There are no mental states of the machine at stake, only very sophisticated model architectures based on probability and trained on huge data sets.

So, when we talk about deception, we are talking about characteristics of the chatbots that can deceive, without assuming that the chatbots *want* or *intend* to deceive.

There is a great debate about whether we can call it manipulation or deception when they have no intention to deceive. At this point, a brief overview of the concepts of manipulation and deception is necessary.

In contrast to deception, which is limited to the epistemic level, manipulation retains a certain semantic connection to the psychological level, namely the idea of being steered in a certain direction (Cohen, 2023). The manipulator’s will to achieve a certain result is

an important aspect of manipulation. Manipulation is action-guiding. Deception can, of course, be used for manipulation, but it remains at the level of belief. According to our definition, deception is the inducement of false beliefs. However, we exclude errors and mistakes that can lead to the formation of false beliefs from our analysis. In the literature on manipulation and deception with technology, it is instructive that the concept of *manipulation* is used when microtargeting is discussed, while the concept of *deception* is used in the case of interaction with social robots. The literature on microtargeting is not about arguing whether recommendation systems have intentions or not, but rather about attributing intentionality back to the humans involved and arguing that technology can be an aggravating factor (Jongepier and Klenk, 2022). In the social robotics literature, there are some efforts to change the definition of deception so that it does not require intentionality. For example, if a human interacting with a robot forms the false belief that the robot has emotions, we can say that the robot has deceived the human, even if the robot has no intentions. An interesting approach from the literature is the notion of banal deception (Natale, 2021), which acknowledges that all media are deceptive, but not in the classical sense of 'deliberate deception', but in a more functional sense that uses tactics to form false beliefs for a better experience with technology, be it a laptop, a social robot, or a voice assistant. Deception is inherent in media, but it is not a form of outright manipulation and it is instrumentally valuable[3].

We believe that we can extend the concept of banal deception to LLM-based chatbots. They exhibit some deceptive features, starting with the most notable ones: anthropomorphism and mimicry. LLM-based chatbots produce outcomes in natural language. This means that users can easily communicate without having to program. Users speak *as if* they were talking to a human, and the chatbot responds *as if* it were a human, i.e., it mimics the human way of speaking or writing. One way to do this is to start the sentences with mentalistic propositions like 'I believe,' 'I want,' or propositions that express an emotion, such as 'I'm sorry.' This has clear advantages because it is more user-friendly, but also some consequences. The most relevant is anthropomorphism: users project onto chatbots mental states they do not have, such as beliefs, intentions, and desires. Dennett's (1987) intentional stance can be very helpful in this respect. Humans transfer their intentions to others in order to understand and predict their

[3] The distinction between banal deception and manipulation is only roughly outlined here; we merely introduced a distinction between manipulation as action guiding and deception as epistemic distortion. It is acknowledged that there may be multiple layers and forms of overlap, and that there may be varied interpretations of manipulation being goal-oriented. A broad interpretation of the distinction could conceive human-likeness as an engagement lever aimed at engagement maximisation, and consequently regard this form of deception as a manipulative device. A narrow interpretation has the potential to exclude such general aims, with the focus instead being placed on specific goals, such as convincing an individual to perform a particular action. However, due to limitations in the available space, a more thorough exploration of this topic is not feasible.

behaviour, even if they do not have access to the mental states of other humans or animals.

The phenomena described by anthropomorphism and the intentional stance are not new, and it is relatively easy for humans to attribute these properties to systems that are not even sophisticated. The fact that LLM-based chatbots are now so sophisticated that they give the impression that 'they understand you' (although sometimes they do not, which is very frustrating) makes it even easier for users to automatically attribute mentalistic states and emotional abilities to them. Another important feature is the rhetorical style of these chatbots. They tend to answer assertively and sometimes in an overly self-assured or even patronising manner. They never seem to have doubts or insecurities (which they do not have either, but they also do not have a certainty or self-confidence that allows for self-assurance). Empirical studies are starting to emerge, although they are still mostly at the under-review or preprint stages. Preliminary findings suggest that anthropomorphism – especially when it showcases intelligence and expertise – can foster trust (Colombatto et al., 2025). However, empirical research highlights how anthropomorphism does not have a clear-cut path towards trust, for example, human-likeness can be not-determinant when not functional to the users' goals (Bouyzourn and Birch, 2025; Haresamudram et al., 2025) or it can improve connection but reduce trust when perceived as non-authentic or non-reliable, or non-credible (Cohn et al., 2024; Basoah et al., 2025; Wang et al., 2025), or it can also evoke unsettling emotions when coupled with hallucinatory or erratic outputs (Rapp et al., 2025). Importantly, perceived self-confidence plays an important role, as it seems humans tend to perceive AI as more self-assured than humans, even when they have an identical performance, because of a prior belief that AI is more accurate (Colombatto et al., 2025). Yet this projected assurance, either perceived or expressed through language, does not actually correspond to accuracy in responses.

Moreover, there are also issues at the interface level: as mentioned in the previous section, OpenAI's ChatGPT does not display a banner explaining what it is and briefly describing how to set our expectations. It merely warns that ChatGPT can make 'mistakes,' fuelling the false belief that its default outputs are 'correct' or 'true.'

These features make it clear that there is a dissonance between what the chatbot appears to be able to do and what it actually does. What interests us is the relationship between deception and trust in LLM-based chatbots.

In the first section we argued that the flaw in trusting LLM-based chatbots lies in the erroneous application of the same trust granted to previous technological systems without taking into account the substantial shift from linear to non-linear AI systems. What these systems appear to be and what they are play a major role in this mismatch. In the first section, this idea was articulated primarily in terms of them appearing to be 'truth-tellers' rather than 'storytellers.' However, we can go further: given the additional

deceptive properties discussed above, we would like to argue that these properties may have additional implications for how they influence the evaluative act of trust.

However, there are different ways to interact with LLM-based chatbots, whereby we can distinguish three main types:

1. *Interaction with chatbots*: When the style of language makes us believe that the machine has mental states, we tend to employ the same structure that we use to evaluate human trustworthiness (anthropomorphism). Here, the discursive tone can be important when it comes to assessing the reliability of the target: the more confidence we have, the more we assume we can trust the target. If the information is false but communicated confidently and we have no prior knowledge, there seems to be a pressure not to exercise our epistemic agency and check. And when it comes to revision of beliefs, chatbots tend to agree with users when the latter express dissatisfaction (what is dubbed as *sycophancy*), reducing the likelihood of revision.
2. *Interaction with chatbots integrated into proprietary apps or websites:* This case is similar to the one above, but is complicated by the additional trust in the brand itself. For example, if we believe that Amazon is a reliable service, it is possible that we transfer trust to the Amazon chatbot.
3. *Interaction with social media bots:* This latter case presents an additional problem, namely that of indistinguishability. If the chatbot is disguised as a personal profile and interacts like one, we are inclined to treat it like a human. However, if we know that these bots are difficult to recognise, we begin to question whether or not we are interacting with a person, or we doubt that we are really interacting with a person. This doubt is not so easy to dispel and can lead us to go the other way round and no longer trust what we see online and disengage from online communication and interaction altogether.

Case 1 seems to be a case of banal deception. Cases 2 and 3 utilise the same psychological mechanisms as in case 1, but can lead to manipulation: deception – in this case, human-likeness – can be used as a means to steer people in a certain direction. In the case of chatbots used in apps, this can be a way to persuade people to make further purchases. In the case of social media bots, they can be used to manipulate users' voting preferences.

Overall, the evaluative act of trust towards LLM-based chatbots is more similar to how we would trust another person than how we would trust a pocket calculator. This is because in the evaluation process we consider not only technical aspects, but also interactive aspects shaped by natural language (without taking into account that some

even report having formed a bond with the chatbot), we react to anthropomorphic cues that users recognise when interacting with chatbots.

This results in a paradox: on the one hand, as just mentioned, the evaluative act of trusting generative AI systems is influenced by the anthropomorphic features of chatbots, such as self-confident tone; on the other hand, the alleged reliability of these systems is inherited from our reliance on digital technologies, just because generative AIs are digital technologies. These two settings create a sort of hybrid target for trust that is both human-like (in the way we interact with chatbots) and artificial (in the way we consider chatbots as linear technologies and therefore reliable and infallible). We are dealing with quasi-subjects that we trust with a type of trust that we would place in human conversational agents, but expecting greater efficiency and reliability than we would expect from human conversational agents.

Now, vulnerability can be seen as a condition of lower levels of epistemic agency: when we have less control, we are more vulnerable. Taken together, these factors explain why interacting with LLM-based systems can make users particularly susceptible to epistemic vulnerability. However, we can distinguish between vertical vulnerabilities and horizontal vulnerabilities. The first refers to groups of people who are considered less epistemically equipped than the average. In the context of technology, children and the elderly are considered more vulnerable groups. However, with the emergence of LLM-based chatbots, a vulnerability has crystallised as a result of the sheer power and widespread use of this technology, and it is a horizontal one: anyone can be vulnerable at this stage of transition. In this context, digital humanism can be a good ally to get through this transitional phase with *more*, rather than less, epistemic agency.

4 Reframing Trust through Digital Humanism

As we have seen, the interactive nature of LLMs creates conditions in which users are particularly susceptible to false beliefs – often without realising it. This dynamic of deception results from design features that simulate human-like communication while concealing the actual limitations of the system. Rather than dismissing these systems outright, we argue that such interactions can serve as a critical lens through which we can re-examine the broader models of trust that we have extended to digital technologies in the past.

Recognising that LLMs are not reliable producers of truth forces us to confront the assumptions embedded in our previous reliance on media technologies. In many cases, these systems have been treated not merely as mediators of information, but as guarantors of epistemic authority. The dissonance introduced by LLMs helps to reveal the fragility of this assumption.

This realisation opens up the space for a broader reconfiguration of our relationship with technology. Trust in digital systems needs to be reframed as a situated and evaluative stance rather than a passive expectation. LLMs, by highlighting the gap between discursive coherence and epistemic accountability, can become cultural artefacts through which we reconsider how information is framed, believed, and acted upon.

Such a reorientation is in line with the ethical and epistemological priorities of digital humanism, which asks us to consider technologies not as neutral tools, but as embedded cultural forms that shape and reflect human values, behaviours, and vulnerabilities.

In current debates about digital humanism, two complementary but methodologically distinct approaches have emerged, both of which offer valuable insights for rethinking the nature of trust in technological environments. Although these approaches originate from different premises, they often converge on the need to reclaim a space for human agency, autonomy, and ethical reflection in the face of technological transformation.

The first approach, associated with initiatives such as the *Vienna Manifesto on Digital Humanism* (Werthner et al., 2022, pp. xi-xiv) and the DigHum network, is primarily normative in its orientation. It begins with the formulation of a set of human values – dignity, freedom, responsibility, justice – and then attempts to translate these values into principles for the design, governance, and evaluation of digital technologies. This model, which we might call *top-down digital humanism*, views technologies not as neutral tools, but as systems whose structure, use, and social embeddedness must be normatively evaluated. It emphasises the importance of public accountability, democratic control, and anticipatory ethical reflection in the development and deployment of digital infrastructures.

In contrast, a second approach, developed in the French tradition of *humanisme numérique*, pursues a *bottom-up hermeneutic method* inspired by the philological and historical practices of Renaissance humanism and the Vichian imperative to let doctrines emerge from the objects they study[4]. This orientation was first systematically articulated by Milad Doueihi (2011), who understood the digital not only as a technical substrate, but as a cultural transformation. Here, technologies are analysed as *cultural artefacts* that are an expression of a historical moment and a particular way of shaping the human. The focus is less on prescribing values from above, but rather on observing how digital artefacts – such as algorithms, interfaces, or LLMs – participate in shaping practices, discourses, and perceptions. In this view, ethical insights emerge from a careful reading of the ways in which technologies transform human life, language, and thought.

[4] See the Position Paper For a Critical Digital Humanism, https://www.yumpu.com/fr/document/read/68683985/2024-mai-positionpaper-hn-en&sa=D&source=docs&ust=1749135910609674&usg=AOvVaw29dm_rbGiJ80dVRm33MGjM.

Despite their methodological divergence, these two approaches are not antagonistic. In fact, they often converge on key issues – most notably the need for inter- and transdisciplinary collaboration, the recognition of the cultural embeddedness of technology, and the prioritisation of human well-being and the common good as central goals of digital transformation. Their complementary perspectives provide a robust framework for analysing technologies such as LLMs that resist simple categorisation as tools.

From a digital humanist point of view – whether top-down or bottom-up – LLMs must be understood not only as technological artefacts, but as cultural products. Or rather, as *techno-cultural artefacts* whose technical construction is inextricably linked to the cultural logics, epistemologies, and power structures they encode. This also entails a critical analysis of power relations inscribed in data production, model training, and platform governance, since the epistemic and economic asymmetries that underpin these domains shape who can speak, be heard, and be trusted in digital spaces. Trained on vast corpora of texts from different domains – filtered, pre-processed, and structured according to specific assumptions about language, knowledge, and relevance – such systems inevitably reflect and reproduce the values, biases, and exclusions inherent in the data they ingest and in the design choices of their creators.

Examining LLMs through a humanistic lens allows us to grasp their normative impact beyond their immediate functionality. These systems influence how knowledge is accessed, how authority is perceived, and how beliefs are formed and stabilised in digital environments. A digital humanist reading of LLMs draws attention to the aesthetic and rhetorical strategies with which these models construct coherence and simulate competence. It also prompts us to ask how such strategies affect users' sense of epistemic agency, autonomy, and interpretive responsibility.

In addition, humanistic disciplines such as philosophy, literary theory, history, and media studies provide tools to situate LLMs in a longer genealogy of knowledge mediation – from the invention of writing to the printing press, from encyclopaedias to search engines. They also encourage a critical examination of power structures and show how technological systems participate in the reproduction or contestation of institutional and discursive hegemonies.

This bottom-up orientation, which pays attention to the symbolic and cultural dimensions of technology, does not exclude a normative critique. On the contrary, it enables a form of *situated normativity* that is grounded in the lived experience of users and the concrete affordances of specific systems. In this respect, the hermeneutic approach converges with the more declarative ethics of the top-down model, especially when it comes to shared goals: the protection of human dignity, the prevention of epistemic harm, and the promotion of environments that favour critical thinking, deliberation, and meaningful participation.

In the context of LLMs, this convergence is particularly fruitful. These models challenge not only our understanding of language and meaning, but also our understanding of trust and knowledge. By analysing LLMs as cultural artefacts that speak in our language, mimic our rhetorical patterns, and mirror our cognitive biases (Vallor, 2024, pp. 48-49), digital humanism equips us with a vocabulary and methodology to critically engage with their implications. It helps us to move beyond the binary of uncritical acceptance or technophobic rejection to a mode of reflection that recognises both the risks and the heuristic value of these technologies.

Ultimately, both approaches to digital humanism call for a renewed cultural literacy – one that includes the ability to read technologies, decode their implicit assumptions, and articulate alternative imaginaries. In this sense, the study of LLMs becomes not only a technical endeavour, but also a philosophical and political one: an invitation to redefine what it means to understand, believe, and trust in the digital age.

A digital humanist approach to trust begins by reframing trust itself – not as a passive expectation or functional reliance, but as an evaluative act, a normative judgement about the trustworthiness of a system in a given context. As argued earlier, such a judgement presupposes an active investment capable of shaping or reinforcing a disposition of confidence. From this perspective, we need to go beyond an instrumental view of technology (which considers reliability as a property related to the success of a linear interaction) and question its systemic role within cultural, social, and political structures. Technologies need to be evaluated not only in terms of their functionality, but also in terms of how they organise interactions, distribute agency, and reproduce or challenge existing asymmetries.

This shift also requires a renewed commitment to digital literacy, understood not just as a set of technical skills but as a capacity for critical orientation. And digital literacy, in a digital humanist sense, cannot be reduced to a demand for transparency. As recent debates in ethics of technology have shown (Alloa, 2022; Alloa and Thomä, 2022; Carbone and Lingua, 2023), transparency often risks becoming a moral and political fetish, as an ideal of total informational openness that paradoxically obscures rather than clarifies the processes it seeks to reveal. Following Striano (2024b), what we need are not merely ‘transparent,’ but honest technologies: systems that make mediation perceptible and negotiable, rather than hidden behind the illusion of full visibility. A literacy grounded in honesty rather than transparency would cultivate interpretive awareness and civic responsibility, encouraging users and institutions alike to engage critically with the limits, biases, and opacities inherent in technological mediation.

However, digital literacy should not be conceived merely as an individual competence. While we argue for cultural literacy, we are aware of power asymmetries between users and companies providing the technological services; it is simply not a level playing field. To ask that only users be responsible would mean denying this reality. Hence, within a

digital humanist approach, we recognise cultural literacy also requires collective infrastructures of accountability and education, as well as public policies that foster critical engagement with AI systems. Strengthening individual epistemic agency must go hand in hand with institutional and civic responsibility.

In practical terms, such a literacy could take shape through interdisciplinary education that combines humanities and computer science, participatory design processes that include users in evaluating algorithmic affordances, and civic initiatives that promote the public understanding of mediation rather than the illusion of transparency. Digital humanism, in this sense, calls for an ecosystem of practices that cultivate not only technical skills but interpretive, ethical, and political sensibilities – an education for reading, designing, and governing technologies as cultural forms.

5 Conclusion

This paper aimed to explore how technology, especially generative AI systems, invites – even forces us – to rethink fundamental concepts such as trust. In this paper, we have argued for a more considered and aware way of interacting with AI systems – one that is guided by the principles of digital humanism.

In the first section, we looked at how trust, which has traditionally been extended to technologies that are deterministic and predictable, has been erroneously applied to generative AI systems, even though they are nonlinear. We follow Shionoya (2001) in arguing that trust should be understood as an evaluative act in which a trustor judges whether another, human or not, is trustworthy under certain conditions. This perspective treats trust not as a fixed property, but as a dynamic, normative practice that involves active judgement. According to this perspective, trust is based on the confidence of the trustor – an underlying disposition or willingness to trust. It allows individuals to ascribe trustworthiness to entities, including non-human systems, not because these systems inherently possess moral qualities, but because they consistently exhibit reliable behaviour. However, trust in artificial systems, such as generative AI, should be based on an evaluation of their peculiar performance and not on the mistaken assumption that they function like deterministic systems.

In the second section, we delved into the relational aspect of trust, focusing on how the human-like qualities of generative AI can be deceptive. We explored how, when interacting with AI, we tend to attribute mental states and intentions to these systems even though they do not have them. This tendency, combined with design choices that make AI sound self-confident or even ‘human,’ often leads us to blindly trust the technology and reduce our ability to critically evaluate the information it provides. As LLM-based chatbots are integrated into almost every app and social media, the risks of

deceptive design can extend to risks of manipulation, whether for the purpose of increasing sales or changing voting preferences.

In the third section, we turned to digital humanism as a conceptual framework for thinking about how we should approach AI. Rather than seeing AI merely as a tool, a human-like subject, or some sort of infallible superhuman intelligence, digital humanism asks us to consider it as part of a larger cultural and ethical landscape. This perspective encourages us to engage *with* these technologies not only functionally, but also critically, and to understand how they shape our values, our behaviour, and our ways of knowing.

Finally, we have emphasised the importance of philosophical enquiry to help us address the challenges that generative AI brings. In conclusion, rethinking trust through a digital humanist lens is a crucial step towards a more critical, ethically responsible, and socially engaged approach to technology.

References

Alloa, E. (ed.), This Obscure Thing Called Transparency: Politics and Aesthetics of a Contemporary Metaphor, Leuven University Press, Leuven 2022.

Alloa, E. and Thomä, D. (eds.) (2022), Transparency, Society and Subjectivity: Critical Perspectives, Palgrave Macmillan, London.

Baier, A. (1986), Trust and Antitrust, in 'Ethics', 96(2), pp. 231-260

Blackburn, S. (2010), Trust, cooperation, and human psychology, in Id. Practical Tortoise Raising and other philosophical essays, Oxford Academic, Oxford, pp. 90-108. https://doi.org/10.1093/acprof:oso/9780199548057.003.0006.

Basoah, J., Chechelnitsky, D., Long, T., Reinecke, K., Zerva, C., Zhou, K., Díaz, M. and Sap, M. (2025), Not Like Us, Hunty: Measuring Perceptions and Behavioral Effects of Minoritized Anthropomorphic Cues in LLMs, in 'Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency', pp. 710-745. https://doi.org/10.1145/3715275.3732045.

Bouyzourn, K. and Birch, A. (2025). What Shapes User Trust in ChatGPT? A Mixed-Methods Study of User Attributes, Trust Dimensions, Task Context, and Societal Perceptions among University Students, in 'ArXiv'. https://arxiv.org/abs/2507.05046v1.

Calì, C. (2023), Come ci cambia la tecnologia. L'Agency delle AI e la capacità cognitiva di prendere decisioni razionali, in 'S&F_scienzaefilosofia.it', 30, pp. 366-385.

Carbone, M. and Lingua, G. (2023), Toward an Anthropology of Screens: Showing and Hiding, Exposing and Protecting, Palgrave Macmillan, London.

Cohen, S. (2023), Are All Deceptions Manipulative or All Manipulations Deceptive?, in 'Journal of Ethics and Social Philosophy', 25(2). https://doi.org/10.26556/jesp.v25i2.1998.

Cohn, M., Mahima Pushkarna, Olanubi, G. O., Moran, J. M., Padgett, D., Mengesha, Z. and Heldreth, C. (2024), Believing Anthropomorphism: Examining the Role of Anthropomorphic Cues on Trust in Large Language Models, in 'Extended Abstracts of the CHI Conference on Human Factors in Computing Systems', pp. 1-15. https://doi.org/10.1145/3613905.3650818.

Colombatto, C., Birch, J. and Fleming, S. M. (2025), The influence of mental state attributions on trust in large language models, in 'Communications Psychology', 3(1). https://doi.org/10.1038/s44271-025-00262-1.

De Fine Licht, K. and Brülde, B. (2021), On defining 'Reliance' and 'Trust': purposes, conditions of adequacy, and new definitions, in 'Philosophia', 49(5), pp. 1981–2001. https://doi.org/10.1007/s11406-021-00339-1.

Deley, T. and Dubois, E. (2020), Assessing trust versus reliance for technology platforms by systematic literature review, in ‘Social Media + Society’, 6(2), pp. 1-8. https://doi.org/10.1177/2056305120913883.

Dennett, D. C. (1987), The intentional stance, The MIT Press, Cambridge (MA).

Département Humanisme Numérique – Collège des Bernardins, For a Critical Digital Humanism (Position Paper), https://www.yumpu.com/fr/document/read/68683985/2024-mai-positionpaper-hn-en&sa=D&source=docs&ust=1749135910609674&usg=AOvVaw29dm_rbGiJ80dVRm33MGjM.

Doueihi, M. (2011), Pour un humanisme numérique, Seuil, Paris.

Edwards, P. N. (2010), A Vast Machine. Computer Models, Climate Data, and the Politics of Global Warming, The MIT Press, Cambridge (MA).

Floridi, L. (2023), AI as Agency Without Intelligence: on ChatGPT, Large Language Models, and Other Generative Models, in ‘Philosophy & Technology’, 36(1). https://doi.org/10.1007/s13347-023-00621-y.

Gorrieri, L. (2024), Is ChatGPT full of bullshit?, in ‘Journal of Ethics and Emerging Technologies’, 34(1), pp. 1-16. https://doi.org/10.55613/jeet.v34i1.149.

Haresamudram, K., Van As, N. and Larsson, S. (2025), Tasks Over Traits: User Perception of Humanlike Features in Goal-Oriented Chatbots, in ‘International Journal of Human-Computer Interaction’, 41(21) pp. 13363–13381. https://doi.org/10.1080/10447318.2025.2470311.

Himma, K.E. (2008), Artificial agency, consciousness, and the criteria for moral agency: what properties must an artificial agent have to be a moral agent?, in ‘Ethics and Information Technology’, 11(1), pp. 19-29. https://doi.org/10.1007/s10676-008-9167-5.

Hinton, G., Bengio, Y., Hassabis, D., Altman, S., Amodei, D., Song, D., Lieu, T., Gates, B., Zhang, Y.Q., Sutskever, I., et al. (2023, May 30), Statement on AI risk [open letter]. https://safe.ai/work/statement-on-ai-risk.

Jongepier, F. and Klenk, M. (2022), Online manipulation. Charting the field, in Jongepier, F. and Klenk, M. (eds.), The Philosophy of Online Manipulation (pp. 15-48), Routledge, New York. https://doi.org/10.4324/9781003205425-3.

Latour, B. (1987), Science in Action: How to Follow Scientists and Engineers Through Society, Harvard University Press, Cambridge (MA).

Rapp, A., Di Lodovico, C., and Di Caro, L. (2025), How do people react to ChatGPT's unpredictable behavior? Anthropomorphism, uncanniness, and fear of AI: A qualitative study on individuals' perceptions and understandings of LLMs' nonsensical hallucinations, in 'International Journal of Human-Computer Studies', 198. https://doi.org/10.1016/j.ijhcs.2025.103471.

Schlosser, M. (2019). Agency, In Zalta, E. (ed.), The Stanford Encyclopedia of Philosophy. Stanford University. Retrieved 13 Apr 2025, from https://www.plato.stanford.edu/archives/win2019/entries/agency/.

Shionoya, Y. (2001), Trust as a Virtue, in Shionoya, Y. and Yagi, K., Competition, Trust, and Cooperation: A Comparative Study, Springer, Berlin-Heidelberg, pp. 3-19.

Star, S. L. (1999), The Ethnography of Infrastructure, in 'American Behavioral Scientist', 43(3), pp. 377-391.

Striano, F. (2024a), Can artificial agents act? Conceptual constellation for a de-humanised theory of action, in 'S&F_scienzaefilosofia.it', 31, pp. 224-244.

Striano, F. (2024b), The Vice of Transparency. A Virtue Ethics Account of Trust in Technology, in 'Lessico di Etica Pubblica', 1/2024, pp. 70-86.

Swanepoel, D. (2021), Does artificial intelligence have agency?, in Clowes, R.W., Gärtner, K., and Hipólito, I., The Mind-Technology Problem, Springer, Berlin-Heidelberg, pp. 83-104. https://doi.org/10.1007/978-3-030-72644-7_4.

Taddeo, M. (2010), Modelling trust in artificial agents, a first step toward the analysis of e-Trust, in 'Minds and Machines', 20(2), pp. 243-257. https://doi.org/10.1007/s11023-010-9201-3.

Taddeo, M. (2017), Trusting digital technologies correctly, in 'Minds and Machines', 27(4), pp. 565-568. https://doi.org/10.1007/s11023-017-9450-5.

Thompson, C. (2018), Faire confiance aux artéfacts – Faire confiance à distance, in Doueihi, M. and Domenicucci, J. (eds.), La confiance à l'ère numérique, Éditions rue d'Ulm, Paris, pp. 97-111.

Vallor, S. (2024), The AI Mirror. How to Reclaim Our Humanity in an Age of Machine Thinking, Oxford University Press, Oxford.

Wang, K., Quek, B.-K., Goh, J. and Herremans, D. (2025), To Embody or Not: The Effect Of Embodiment On User Perception Of LLM-based Conversational Agents, in 'ArXiv'. https://doi.org/10.48550/arXiv.2506.02514.

Werthner, H. et al. (2022), Vienna Manifesto on Digital Humanism, in Werthner, H., Prem, E., Lee, E. A., Ghezzi, C. (eds.), Perspectives on digital humanism, Springer Cham, pp. xi-xiv

AI in a Class-Diverse India: Rights, Representation, and Regulation

Soumya Singh Chauhan

Jindal Global University, Sonipat, India

DOI 10.3217/978-3-99161-062-5-008, CC BY 4.0

Abstract. The integration of Artificial Intelligence (AI) into governance frameworks is accelerating across the Global South, and India stands at the forefront of this transformation. From biometric welfare systems and predictive policing to algorithmic surveillance, AI is increasingly embedded in public service delivery and state infrastructure. However, this technological expansion occurs within a socio-political landscape deeply shaped by caste, religion, and economic class. This paper critically interrogates how AI systems intersect with India's entrenched hierarchies, revealing the representational, regulatory, and ethical gaps that threaten to reproduce and entrench structural injustice.

Drawing from interdisciplinary frameworks in AI ethics, critical data studies, and postcolonial science and technology studies, the paper engages with concepts such as sociotechnical imaginaries, algorithmic discrimination, and data colonialism. It explores how digital systems often erase class-based identities, resulting in opaque decision-making, discriminatory surveillance, and the erosion of privacy and agency for marginalized communities. Through case studies of facial recognition, welfare exclusion, and predictive policing, the paper demonstrates how caste, religious, and economic markers are indirectly encoded into algorithmic governance.

The analysis reveals that India's techno-solutionist regulatory model prioritizes innovation and efficiency over rights, accountability, and inclusion. The Digital Personal Data Protection Act, 2023, fails to address algorithmic discrimination, ensure transparency, or mandate oversight. In response, the paper proposes a rights-based, class-conscious AI governance model rooted in India's constitutional commitments to equality, justice, and fraternity. It calls for participatory design, disaggregated data practices, and robust accountability mechanisms to ensure AI serves as a tool of inclusion rather than oppression.

1 Introduction

Artificial Intelligence (AI) is rapidly transforming the landscape of governance and public service delivery in India, mirroring the global phasing in of the same. From biometric identification systems to algorithmic credit scoring and AI-driven surveillance, the state is increasingly embedding AI technologies into the architecture of administration and control.[1] This integration, however, is not occurring in a social vacuum. It is unfolding within a deeply stratified society, where caste, religion, gender, and class shape not only individual lives but also institutional logics and state practices.[2] [3]

India's socio-political fabric is characterized by vast and persistent inequalities. The intersection of caste-based exclusion, religious discrimination, and economic marginalization continues to structure access to rights, recognition, and resources.[4] In such a context, the presumed neutrality of AI technologies—frequently celebrated for their efficiency and objectivity—must be interrogated. AI systems do not merely automate decisions; they encode values, histories, and exclusions.[5] When deployed without attention to the complexities of Indian society, they risk reproducing and even amplifying existing hierarchies under the guise of modernization.[6]

This paper asks three central research questions:

---

[1] Eubanks, V. (2018). Automating inequality: How high-tech tools profile, police, and punish the poor. St. Martin's Press.

[2] Teltumbde, A. (2018). *Republic of caste: Thinking equality in the time of neoliberal Hindutva*. Navayana.

[3] **Positionality Statement:**

The author acknowledges her social position outside the lived experiences of caste-based and religion-based discrimination, algorithmic violence, and surveillance targeting marginalized communities. This paper engages with questions of caste, data disaggregation, surveillance, and digital exclusion by drawing upon the work of Dalit, Bahujan, Adivasi, Muslim, and working-class scholars, activists, and civil liberties organizations. The author does not claim epistemic authority over these perspectives, nor can she fully convey the affective and material weight of being profiled, misrecognized, or erased by technological systems. Calls for caste-conscious data governance, rights-based regulation, and participatory frameworks are made here with humility, and with the recognition that even well-intentioned interventions risk reproducing extractive logics if they are not guided by those most affected. Discussions of disaggregated data, algorithmic representation, and techno-legal reform must centre the voices and leadership of communities historically excluded from knowledge production and decision-making. This paper therefore positions itself as a contribution to and not as a substitute for broader, community-led critiques of epistemic and infrastructural injustice in AI governance in India. It is an invitation to continued engagement, correction, and collaborative transformation.

[4] Jaffrelot, C. (2021). *Modi's India: Hindu nationalism and the rise of ethnic democracy*. Princeton University Press.

[5] Benjamin, R. (2019). *Race after technology: Abolitionist tools for the new Jim code*. Polity.

[6] Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism*. NYU Press.

1. How do AI systems interact with and potentially reinforce class-based hierarchies in India?
2. What representational and regulatory gaps exist that allow for the perpetuation of bias and exclusion through AI technologies?
3. What would an equitable and context-sensitive model of AI governance look like in a society as structurally unequal and diverse as India?

To answer these questions, the paper adopts an interdisciplinary approach that draws from legal studies, critical data science, and postcolonial science and technology studies. It engages with conceptual frameworks such as sociotechnical imaginaries, algorithmic discrimination, and data colonialism, and situates them within India's historical and contemporary structures of inequality. The analysis weaves together case studies, policy critique, and normative frameworks to assess the risks of uncritical AI adoption in state functions.

The paper is structured as follows: it begins with the theoretical framework that informs the critique, followed by an exploration of India's socio-historical context of inequality. It then analyses how AI systems reproduce social disparities, challenges the myth of technological neutrality, and evaluates the surveillance architecture and its disproportionate effects on marginalized communities. The regulatory and legal landscape is then examined, highlighting both domestic gaps and international best practices. Finally, the paper outlines a rights-based, inclusive vision for AI governance in India, one that centres justice, representation, and structural reform over technocratic efficiency.

2 Theoretical Framework

To critically evaluate the role of AI in reproducing class-based hierarchies in India, this paper draws from interdisciplinary frameworks in AI ethics, critical data studies, and postcolonial science and technology studies. These perspectives reveal how ostensibly neutral technologies are socially and politically embedded and often reinforce historical structures of inequality.

One key concept is sociotechnical imaginaries, which refer to collectively held visions of the future that are enacted through technological development.[7] In India, these imaginaries are often shaped by aspirations of 'smart governance,' 'Digital India,' and 'technological sovereignty.' AI becomes a symbol of progress and modernity, even when

[7] Jasanoff, S., & Kim, S.-H. (2009). *Containing the atom: Sociotechnical imaginaries and nuclear power in the United States and South Korea. Minerva*, 47(2), 119–146. https://doi.org/10.1007/s11024-009-9124-4

implemented without adequate attention to justice, consent, or social difference.[8] These imaginaries obscure the fact that technologies are not neutral tools, but instruments embedded in political and institutional contexts.

The concept of data colonialism provides a useful lens to critique the extractive logics of AI in postcolonial societies. Even when developed domestically, AI systems in India are often built on paradigms that commodify human life through digital data.[9] The absence of community consent, the aggregation of behavioural and biometric data without oversight, and the transnational flow of such data to private or state actors mirror colonial forms of dispossession, only now digitized.

A related concept is algorithmic bias, which refers to the ways in which AI systems inherit and magnify social inequalities encoded in historical data.[10] In the Indian context, where caste, religion, and economic status shape access to services and opportunities, datasets often reflect these structural disparities. Yet AI systems built on such data are presented as objective or efficient, hiding their potential for exclusion.

Finally, the notion of surveillance capitalism helps explain the state's increasing reliance on AI technologies for monitoring, profiling, and regulating populations.[11] These systems extract behavioural data to classify individuals into risk categories, often without transparency or accountability. When used by the state, such surveillance is not only a matter of privacy but of political control, with disproportionate impacts on already marginalized groups.

Together, these concepts challenge the assumption that AI technologies can be separated from the societies in which they are developed and deployed. In a nation marked by enduring hierarchies, a class-neutral AI policy is not merely inadequate, it risks becoming an instrument of automated injustice.

---

[8] Ghosh, B., & Arora, S. (2019). *Smart as democratically transformative? An analysis of 'Smart City' sociotechnical imaginary in India.* IDS/Steps Centre Working Paper 109.

[9] Couldry, N., & Mejias, U. A. (2019). *The costs of connection: How data is colonizing human life and appropriating it for capitalism*. Stanford University Press.

[10] Eubanks, V. (2018). *Automating inequality*. St. Martin's Press; Noble, S. U. (2018). *Algorithms of oppression*. NYU Press.

[11] Zuboff, S. (2019). *The age of surveillance capitalism*. PublicAffairs.

3 Literature Review

A growing body of interdisciplinary scholarship has raised critical concerns about the intersection of artificial intelligence (AI), surveillance, and systemic inequality, particularly in societies where legal safeguards are minimal, transparency is lacking, and historic social hierarchies remain entrenched.

Virginia Eubanks (2018) offers one of the most influential accounts in this area. In *Automating Inequality*, she demonstrates how algorithmic decision-making in welfare systems reproduces poverty and discrimination, disproportionately harming low-income, racialized, and otherwise marginalized populations.[12] Kate Crawford (2021), in *Atlas of AI*, extends this critique by tracing how AI systems are embedded in extractive logics, mining not just data, but also human labour, environmental resources, and social hierarchies, ultimately reinforcing global and historical asymmetries of power.[13]

Technical audits by Inioluwa Deborah Raji and Joy Buolamwini (2019) have shown that commercial facial recognition systems exhibit significant accuracy disparities across skin tone and gender.[14] Their research highlights that darker-skinned individuals and women face higher error rates, a risk particularly relevant in India's caste-stratified and class-divided context, where facial recognition is increasingly used in welfare delivery, law enforcement, and exam surveillance. This aligns with findings by Obermeyer et al. (2019), who demonstrated that a healthcare risk algorithm trained on cost-based proxies systematically underestimated the health needs of Black patients in the United States.[15] These insights suggest parallel dangers in India, where predictive analytics in welfare and public health may encode structural exclusions through proxy indicators.

Within the Indian context, Anand Teltumbde (2018) and Suraj Yengde (2019) argue that caste is routinely erased in data collection and yet reappears implicitly through variables such as education, geographic location, and surnames.[16] These proxies are often used in algorithmic decision-making and resource allocation, reinforcing caste hierarchies

---

[12] Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press.

[13] Crawford, K. (2021). *Atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press.

[14] Raji, I. D., & Buolamwini, J. (2019). Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 429–435. https://doi.org/10.1145/3306618.3314244

[15] Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464), 447-453. https://doi.org/10.1126/science.aax2342

[16] Teltumbde, A. (2018). *Republic of caste: Thinking equality in the time of neoliberal Hindutva*. Navayana; Yengde, S. (2019). *Caste matters*. Viking.

without formal recognition of caste. Their work highlights the unique challenge of caste-blind AI systems that replicate social bias under a veneer of neutrality.

At the global level, international human rights frameworks have increasingly sought to place normative limits on the deployment of AI in governance. The *United Nations High Commissioner for Human Rights Report on the Right to Privacy in the Digital Age* (2021) underscores that AI systems must comply with principles of legality, necessity, and proportionality, especially when used in surveillance or law enforcement.[17] These standards serve as critical benchmarks for evaluating India's expanding AI surveillance architecture, which currently lacks strong procedural safeguards, transparency, or independent oversight.

Taken together, this literature reflects a growing consensus that AI governance must be grounded in human rights, social justice, and structural reform, especially in postcolonial democracies like India, where technologies are being introduced into historically unequal infrastructures. The need for disaggregated data, participatory governance, and regulatory frameworks attuned to local contexts is urgent and well established in this emerging field.

4 Social and Historical Context of Inequality in India

To understand the risks posed by AI systems in India, it is necessary to situate them within the country's deeply stratified social order. Caste, religion, and class are not peripheral identity markers but enduring structures that shape institutional access, state power, and socio-economic mobility. AI systems introduced into this terrain do not operate neutrally; rather, they inherit and can reinforce these embedded hierarchies.

Caste remains among the most resilient systems of stratification in India, governing access to education, housing, employment, and justice.[18] Despite constitutional protections and affirmative action, caste-based discrimination persists in both explicit and invisible forms. As Anand Teltumbde and Suraj Yengde argue, neoliberalism has not dismantled caste, it has merely privatized and obscured it.[19] When digital infrastructures ignore caste, they risk reproducing its logic in algorithmic form.

---

[17] United Nations High Commissioner for Human Rights. (2021). *The right to privacy in the digital age* (A/HRC/48/31). https://www.ohchr.org/en/documents/thematic-reports/ahrc4831-right-privacy-digital-age-report-united-nations-high [accessed June 15th, 2024]

[18] Thorat, S., & Neuman, K. (2012). *Blocked by caste: Economic discrimination in modern India*. Oxford University Press.

[19] Teltumbde, A. (2018). *Republic of Caste*; Yengde, S. (2019). *Caste Matters*.

Religion, especially Islam, has become another axis of algorithmic risk. The securitization of Muslim identity through laws, surveillance, and social media monitoring has intensified in recent years.[20] AI tools used in predictive policing or facial recognition often reflect and amplify this bias, especially when trained on data shaped by communal profiling.

Economic inequality intersects with both caste and religion. Despite welfare schemes and digital inclusion initiatives, India's rapid digitization has often increased exclusion at the margins. Errors in biometric authentication (e.g., Aadhaar), lack of mobile access, and opaque algorithmic assessments have disproportionately harmed Dalits, Adivasis, Muslims, and informal workers.[21] These harms are not anomalies, they are systemic outcomes of technologies designed for a universal subject who rarely reflects India's socio-economic majority.

A key challenge in analysing these exclusions lies in the absence of disaggregated data[22]. Most Indian digital governance systems collect minimal or no data on caste, religion, or class, citing neutrality or efficiency. Yet this omission leads to statistical erasure, preventing any meaningful visibility into how AI systems impact marginalized groups.[23] Calls for disaggregation are thus not just technical demands but political claims to recognition.

This socio-historical context makes clear that AI systems in India do not simply fail by accident. When implemented without attention to caste, religion, and class, they succeed in precisely the terms they were designed: to serve the dominant social order while rendering marginality invisible.

---

[20] Jaffrelot, C. (2021). *Modi's India*; Jamil, G. (2017). *Muslim Women Speak: Of Dreams and Shackles*.

[21] Panigrahi, S. (2022). *Marginalized Aadhaar: India's Aadhaar biometric ID and mass surveillance. *ACM Interactions*, 29*(2), 16–22.; Frontline. (2024, December 12). *Mandatory Aadhaar authentication leads to exclusion of the marginalised from PDS.*; The Hindu. (2017, February 18). *Aadhaar no standout performer in welfare delivery.*

[22] Disaggregated Data: data that has been broken down by detailed sub-categories, for example by marginalised group, gender, region or level of education. Disaggregated data can reveal deprivations and inequalities that may not be fully reflected in aggregated data. https://www.right-to-education.org/monitoring/content/glossary-disaggregated-data

[23] Vaidehi, R., Reddy, A. B., & Banerjee, S. (2021). Explaining caste-based digital divide in India. arXiv.

Kumar, A. (2022). Ignoring caste and denying development. Data4SDGs.

5 Bias and Inequity in AI Systems

AI systems in India are increasingly deployed across critical sectors such as welfare, policing, employment, and credit scoring.[24] These systems are often presented as neutral, objective, and scalable, promising efficient governance and rational decision-making. However, when built on biased or incomplete data, AI does not eliminate discrimination; it automates it.[25]

The first and most pressing issue is the absence of disaggregated data. Most AI systems in India do not collect or analyse information based on caste, religion, or socio-economic status. This creates an epistemic gap that conceals how marginalized groups are affected. Facial recognition systems, for instance, may perform poorly on darker-skinned individuals, many of whom are Dalits, Adivasis, or Muslims, but without disaggregated error reporting, this harm remains undocumented.[26] Similarly, hiring algorithms that use proxies like educational background or location may indirectly filter out candidates from historically oppressed communities.[27]

Predictive policing tools trained on biased crime data are another key concern. Studies in India and globally have shown that algorithmic policing tends to over-surveil poor, minority, and politically active populations.[28] In India, this translates into the overrepresentation of Muslims, Dalits, and urban poor as potential threats. Historical policing records, which already reflect decades of communal bias and caste-based targeting, become the foundation for AI systems that perpetuate that bias at scale.

The problem is not only technical but systemic: AI design and deployment in India lack transparency, accountability, and public oversight. There is no legal requirement for algorithmic audits or impact assessments. Civil society actors rarely have access to the

---

[24] Marda, V. (2018). *Artificial Intelligence Policy in India: A Framework for Engaging the Limits of Data-Driven Decision-Making.* Philosophical Transactions A: Mathematical, Physical and Engineering Sciences, Available at SSRN: https://ssrn.com/abstract=3240384 or http://dx.doi.org/10.2139/ssrn.3240384; Khurana, L, et. al. (2025). *Fintech And Financial Inclusion In India: A Data-Driven Analysis Of Digital Payments And Banking Access.* Journal of Informatics Education and Research. Vol 5 Issue 3

[25] Eubanks, V. (2018). *Automating inequality*. St. Martin's Press.

[26] Jain, G., & Parsheera, S. (2021). *Cinderella's shoe won't fit Soundarya: An audit of facial processing tools on Indian faces.* arXiv. https://doi.org/10.48550/arXiv.2112.09326

[27] Benjamin, R. (2019). *Race after technology*. Polity.

[28] Rina Chandran**. (2023).** *India's scaling up of AI could reproduce casteist bias, discrimination against women and minorities.* https://scroll.in/article/1055846/indias-scaling-up-of-ai-could-reproduce-casteist-bias-discrimination-against-women-and-minorities *[accessed June 12th, 2024]*

data or models used in decision-making.[29] The result is a class-blind and caste-unaware AI ecosystem that protects dominant interests while invisibilizing harm.

Moreover, even calls for bias mitigation through disaggregation must be approached critically. Scholars warn that disaggregated data, while important for detecting harm, can also entrench problematic social categories if used without community control or ethical safeguards.[30] Surveillance systems that categorize citizens by caste or religion may end up reinforcing stigma rather than promoting equity.

In sum, AI systems deployed in India today operate within—and often reproduce—inequitable social structures. Without structural reform and inclusive design, these systems risk becoming tools of 'automated inequality.'

6 Rights, Representation, and the Myth of Neutrality

One of the most insidious features of AI systems is the myth of neutrality—the claim that algorithms merely reflect data without political or ethical content. This myth legitimizes a form of technocratic governance that hides systemic exclusion behind a facade of objectivity.[31] In the Indian context, where identities like caste, religion, and socio-economic status shape access to rights and resources, this neutrality is both epistemically and politically violent.

AI systems are often designed without disaggregated representation in training data. In doing so, they commit a form of *epistemic injustice*—the marginalization of certain groups' lived realities and knowledge systems in the very tools meant to serve them.[32] Dalit, Adivasi, Muslim, and working-class communities are rendered invisible in datasets and, by extension, in algorithmic governance. Their needs are neither modelled nor prioritized, leading to exclusion that is both systematic and untraceable.

The absence of representation also impacts the framing of fairness in AI. Fairness metrics, if defined only in mathematical terms, fail to account for the historical and social context of discrimination.[33] For instance, a credit scoring model may treat all defaults equally without recognizing how structural poverty limits financial resilience in oppressed

---

[29] Marda, V. (2020). *Algorithmic accountability in India: A civil society perspective.* Medianama.; Chahal, V, Hooda, S. (2024). *Auditing AI: What is it and why does it matter for India?* Observer Research Foundation.

[30] Noble, S. U. (2018). *Algorithms of oppression*; Benjamin, R. (2019). *Race after technology*.

[31] O'Neil, C. (2016). *Weapons of math destruction: How Big Data Increases Inequality and Threatens Democracy.* Crown Publishing.

[32] Fricker, M. (2007). *Epistemic injustice: Power and the ethics of knowing*. Oxford University Press.

[33] Barocas, S., Hardt, M., & Narayanan, A. (2019). *Fairness and machine learning*. fairmlbook.org.

communities. Similarly, an exam surveillance system may apply the same facial recognition algorithm to all candidates, ignoring how Dalit and tribal students may face misrecognition or digital exclusion due to technical or infrastructural disparities.[34]

Who gets to define fairness, and whose values shape the algorithmic process, are fundamentally political questions.[35] In India, where technological design is dominated by upper-caste, urban, English-speaking actors, the perspectives of those most vulnerable to AI harms are rarely included in development or policy spaces. This asymmetry results not just in misrepresentation but in systemic non-recognition.

There is growing consensus in critical data studies that disaggregated data is essential to identifying and remedying these harms.[36] However, this approach also carries risks. Without safeguards, such data can be co-opted to justify new forms of profiling or surveillance. Scholars caution that disaggregation must not become a technocratic fix to a political problem.[37] It must be paired with community consent, legal protections, and participatory governance mechanisms that ensure such data serves the interests of the communities it describes.

The invisibilization of caste, religion, and class in data is not simply a technical oversight, it is a political act with real-world consequences. Systems built on such erasures deny people the ability to be seen, heard, or served by the technologies that increasingly govern their lives. Confronting this requires more than bias audits or data collection protocols, it demands a rethinking of what it means to design just technologies in a deeply unjust world.

7 Surveillance and Disproportionate Impacts

India has rapidly expanded its use of AI-powered surveillance in the name of administrative efficiency and national security. From facial recognition systems at protests and airports to predictive policing tools in urban centers and politically sensitive regions, AI surveillance is becoming a core apparatus of state power.[38] While marketed

---

[34] Reuters. (2020, November 10). *''Unfair surveillance'? Online exam software sparks global student revolt." Times of India*, reporting on Thomson Reuters Foundation coverage.

[35] Winner, L. (1980). Do artifacts have politics? *Daedalus*, 109(1), 121–136.

[36] Sambasivan, N., Arnesen, E., Hutchinson, B., Doshi, T., & Prabhakaran, V. (2021). *Re-imagining algorithmic fairness in India and beyond.* In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (FAccT '21) (pp. 315–328). Association for Computing Machinery

[37] Benjamin, R. (2019). *Race after technology*. Polity.

[38] Internet Freedom Foundation. (2024, January 16). *Resist Surveillance Tech, Reject Digi Yatra*. Internet Freedom Foundation.

as neutral and technocratic, these systems disproportionately affect Muslims, Dalits, Adivasis, and the urban poor—communities already over-policed and under-protected.[39]

Surveillance systems such as Digi Yatra, AFRS, and the Jarvis prison monitoring platform illustrate the state's growing investment in real-time biometric and behavioural tracking.[40] These tools are often deployed without public consultation, legal transparency, or democratic oversight. In practice, they convert social and spatial disadvantage into algorithmic suspicion. Muslim neighbourhoods become 'high-risk zones'; poor, informal workers become data points for risk scoring.

Predictive policing, in particular, reflects the dangers of algorithmic circularity. Historical crime data, often shaped by caste and communal biases, are fed into machine learning models that then 'predict' future risk in the same communities.[41] The result is not predictive justice but pre-emptive punishment. Innocent individuals are flagged based on where they live, how they look, or what language they speak.

This form of surveillance threatens not just informational privacy but behavioural and decisional privacy, i.e. the freedom to think, act, and move without being watched.[42] When protestors are identified and tracked using facial recognition, or when students are monitored during exams through AI-powered webcams, surveillance becomes a tool of discipline and deterrence.[43] The chilling effect is especially severe for historically marginalized groups, for whom even minor errors in identification can result in disproportionate harm, including arrest, loss of services, or reputational damage.

India's legal framework provides few protections against such overreach. There are no binding transparency norms, audit mandates, or meaningful redress mechanisms for individuals misidentified or wrongly profiled by AI tools. State agencies often invoke national security to avoid scrutiny, citing exemptions in the Digital Personal Data Protection Act (DPDPA), 2023. This creates a governance gap where AI surveillance grows unchecked, especially in spaces of political dissent or social vulnerability.

AI-powered surveillance is not only a question of technology—it is a question of power. In the absence of legal safeguards and public accountability, it becomes a tool of

---

[39] Singh, S., & Mohanty, R. (2023). *Impacts and ethics of using Artificial Intelligence (AI) by the Indian Police. Police Practice and Research*, 24(3), 102–116.

[40] Abhijit Ahaskar (2019). *Uttar Pradesh prisons turn to AI-based video surveillance to monitor inmates.* *https://www.livemint.com/technology/tech-news/uttar-pradesh-prisons-turn-to-ai-based-video-surveillance-to-monitor-inmates-11573196335267.html* *[accessed June 12th, 2024]*

[41] Ramachandran Murugesan (2021). *Predictive policing in India: Deterring crime or discriminating minorities?.* *https://blogs.lse.ac.uk/humanrights/2021/04/16/predictive-policing-in-india-deterring-crime-or-discriminating-minorities/* *[accessed June 12th, 2024]*

[42] Solove, D. (2008). *Understanding privacy*. Harvard University Press.

[43] India Today. (2024). UPSC to deploy AI for exam surveillance.

structural domination. Any ethical AI policy must begin by asking not what can be surveilled, but who is being watched—and why.

8 Regulatory and Legal Gaps

Despite the rapid adoption of AI across state institutions in India, the country's legal and regulatory framework remains profoundly underdeveloped. The Digital Personal Data Protection (DPDP) Act, 2023, India's first comprehensive data protection law, offers limited safeguards against AI-driven discrimination and surveillance.[44] Its focus on consent and individual control over personal data does not address deeper structural harms like algorithmic bias, profiling, or exclusion.

Crucially, the Act grants sweeping exemptions to state actors in matters concerning sovereignty, public order, and national security, effectively insulating state-led AI surveillance from accountability.[45] There are no legal obligations for government agencies to disclose their use of AI systems, conduct impact assessments, or allow public auditing of algorithms.[46] This is particularly concerning given the growing evidence that AI systems deployed in India often reproduce social hierarchies and target already marginalized communities.

Furthermore, the DPDP Act does not mandate disaggregated data collection across caste, religion, gender, or class, nor does it require public agencies to publish impact assessments based on these variables. As a result, algorithmic harms to specific groups remain legally invisible, and therefore unaddressed.[47] There are also no remedies for individuals adversely affected by AI decisions, such as those misidentified by facial recognition or denied services due to algorithmic scoring.

In contrast, international frameworks such as the European Union's General Data Protection Regulation (GDPR) and the EU AI Act provide more robust protections. These include rights to explanation, obligations for transparency in automated decision-making, and mandatory human rights impact assessments for high-risk AI systems.[48] Similarly,

---

[44] *Digital Personal Data Protection Act*, 2023

[45] Krishna Preetham Kanthi. (2024). *Privacy, Surveillance, and State Interest: Appraising the DPDP Act through a Constitutional Perspective. Beyond Encryption: Tech & Data Protection,* https://www.ijlt.in/post/privacy-surveillance-and-state-interest-appraising-the-dpdp-act-through-a-constitutional-perspect *[accessed June 12th, 2024]*

[46] Internet Freedom Foundation. (2023). *DPDP Act analysis: Surveillance and public accountability*.

[47] Sambasivan, N., Arnesen, E., Hutchinson, B., Doshi, T., & Prabhakaran, V. (2021). *Re-imagining algorithmic fairness in India and beyond.* In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. arXiv:2101.09995v2

[48] *Regulation (EU) 2024/1689*

UNESCO's *Recommendation on the Ethics of Artificial Intelligence* (2021) emphasizes fairness, inclusivity, and the right to participate in decisions about AI systems that affect communities.[49]

India's **techno-legal discourse**—that is, the body of policy documents, official strategies, and legal debates surrounding emerging technologies—**remains innovation-driven but accountability-poor**. Government strategies such as *NITI Aayog's National Strategy for Artificial Intelligence* (2018) and *Digital India* emphasize economic growth and 'AI for All,' but devote little attention to human rights, transparency, or oversight mechanisms.[50] Scholars and policy analysts have similarly noted that India's regulatory imagination privileges technological innovation over ethical and legal accountability. The absence of a dedicated AI law, the lack of an independent oversight body, and the government's discretionary power to bypass privacy protections create a governance vacuum. Civil society actors have consistently demanded stronger legal frameworks that address the specific risks posed by AI, including casteist profiling, communal surveillance, and the erasure of minority voices from digital systems.[51] This discourse comprises both *state-led policy narratives*—framing AI and digital governance primarily as engines of national innovation—and *civil society critiques* highlighting the absence of enforceable accountability norms. The tension between these two positions defines India's techno-legal trajectory today.

Regulation cannot merely be reactive or sectoral. It must be proactive, intersectional, and rooted in constitutional values of equality, justice, and fraternity. Without this, AI will continue to operate in India as a class-blind, caste-silent, and surveillance-heavy apparatus of governance.

9 Toward an Inclusive AI Policy

An equitable AI governance framework in India must begin with the recognition that neutrality is not justice. The country's socio-technical systems operate in the shadow of caste, communalism, and economic inequality. Without explicit safeguards, AI technologies will continue to reinforce these asymmetries under the guise of modernization.

---

[49] UNESCO. (2021). *Recommendation on the Ethics of Artificial Intelligence*.

[50] NITI Aayog. (2018). *National Strategy for Artificial Intelligence: #AIforAll.* Government of India. Available at: https://www.niti.gov.in

[51] Gupta, M. (2025). *Regulating Artificial Intelligence in India: A Legal Imperative for Ethical Accountability and Responsible Innovation.* Lawful Legal.; *AI Regulation in India: Between Innovation and Accountability.* (2024). The Policy POV.; Agarwal, A., & Nene, M. J. (2025). *Incorporating AI Incident Reporting into Telecommunications Law and Policy: Insights from India.* arXiv preprint.

The first step toward an inclusive framework is the mandatory collection and use of disaggregated data. AI systems must be able to reflect how their outcomes affect people differently based on caste, religion, gender, and economic status.[52]However, this disaggregation must not be technocratically imposed. It must be co-designed with the communities it aims to represent, governed by data sovereignty principles, and accompanied by ethical safeguards to prevent misuse in surveillance or profiling.[53]

Second, AI systems used in public governance should be subject to mandatory algorithmic audits, especially for high-risk applications like policing, welfare, education, and credit. These audits must include fairness assessments, not just accuracy checks.[54] Audit bodies should be independent, publicly funded, and include diverse representation from civil society, academia, and impacted communities.

Third, India must establish legal protections against algorithmic discrimination, modelled on both international best practices and its own constitutional guarantees under Articles 14, 15, and 21. These protections should include the right to explanation, the right to opt-out of automated decision-making, and remedies for algorithmic harms.[55]

Fourth, a class-aware AI policy must be participatory. This means involving marginalized communities not only as subjects of impact assessments, but as co-creators in design, deployment, and oversight. Grassroots organizations, public interest technologists, and community media must be empowered to critique, shape, and challenge AI systems.[56]

Finally, AI education and policy discourse must move beyond elite institutions and urban centers. Public education campaigns on algorithmic rights, data justice, and digital harm are essential to counter the opacity that currently shields AI systems from scrutiny.[57] Without a broad democratic base, technological governance risks becoming a tool for elite consolidation.

In a class-based democracy, technological design must be accountable to those it most affects. Equity is not an afterthought; it is the measure of legitimacy. India's constitutional values of justice, equality, and fraternity must be at the centre of any AI governance

---

[52] Kitchin, R. (2014). *The data revolution: Big data, open data, data infrastructures and their consequences*. Sage.

[53] Couldry, N., & Mejias, U. A. (2019). *The costs of connection*. Stanford University Press.

[54] Raji, I. D., & Buolamwini, J. (2019). Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products. *Proceedings of the AAAI/ACM Conference on AI Ethics and Society*.

[55] European Commission. (2021). *Proposal for a Regulation on Artificial Intelligence*.

[56] D'Ignazio, C., & Klein, L. F. (2020). *Data feminism*. MIT Press.

[57] Zuboff, S. (2019). *The age of surveillance capitalism*. PublicAffairs.

model. Without this, AI will not be a tool for liberation, but a new instrument of exclusion in digital form.

10 Conclusion

As India deepens its investment in Artificial Intelligence across governance, welfare, and security, it must confront a difficult truth: AI systems, if left unchecked, will not disrupt social hierarchies, they will entrench them. Designed and deployed within a structurally unequal society, these technologies do not merely reflect injustice; they encode, amplify, and automate it.

This paper has argued that the apparent neutrality of AI masks profound representational and regulatory failures. In a context marked by caste stratification, religious marginalization, and economic exclusion, the absence of disaggregated data, legal safeguards, and participatory governance leaves vulnerable communities disproportionately exposed to algorithmic harm.

An equitable AI future in India requires more than technical correction. It demands a structural reckoning. Regulation must be rights-based. Data must be collected with care, consent, and justice in mind. And communities most affected must not be relegated to footnotes, they must be cantered as architects of the systems that govern them.

In a constitutional democracy that promises justice, equality, and dignity for all, technology must be held to those same standards. AI must be accountable not only to efficiency metrics, but to the people, and especially to those whom history has taught to expect neither fairness nor visibility from the state. Only then can Artificial Intelligence become a tool for social transformation, rather than digital domination.

References

Abhijit Ahaskar (2019). Uttar Pradesh prisons turn to AI-based video surveillance to monitor inmates. https://www.livemint.com/technology/tech-news/uttar-pradesh-prisons-turn-to-ai-based-video-surveillance-to-monitor-inmates-11573196335267.html [accessed June 12th, 2024]

Agarwal, A., & Nene, M. J. (2025). *Incorporating AI Incident Reporting into Telecommunications Law and Policy: Insights from India.* arXiv preprint.

AI Regulation in India: Between Innovation and Accountability. (2024). The Policy POV.

Barocas, S., Hardt, M., & Narayanan, A. (2019). Fairness and machine learning. fairmlbook.org.

Benjamin, R. (2019). Race after technology: Abolitionist tools for the new Jim code. Polity.

Chahal, V, Hooda, S. (2024). Auditing AI: What is it and why does it matter for India? Observer Research Foundation.

Couldry, N., & Mejias, U. A. (2019). The costs of connection: How data is colonizing human life and appropriating it for capitalism. Stanford University Press.

Crawford, K. (2021). Atlas of AI: Power, politics, and the planetary costs of artificial intelligence. Yale University Press.

D'Ignazio, C., & Klein, L. F. (2020). Data feminism. MIT Press.

Digital Personal Data Protection Act, 2023

Eubanks, V. (2018). Automating inequality: How high-tech tools profile, police, and punish the poor. St. Martin's Press.

European Commission. (2021). Proposal for a Regulation on Artificial Intelligence.

Fricker, M. (2007). Epistemic injustice: Power and the ethics of knowing. Oxford University Press.

Ghosh, B., & Arora, S. (2019). Smart as democratically transformative? An analysis of 'Smart City' sociotechnical imaginary in India. IDS/Steps Centre Working Paper 109.

Gupta, M. (2025). *Regulating Artificial Intelligence in India: A Legal Imperative for Ethical Accountability and Responsible Innovation.* Lawful Legal.

India Today. (2024). UPSC to deploy AI for exam surveillance.

Internet Freedom Foundation. (2023). DPDP Act analysis: Surveillance and public accountability.

Internet Freedom Foundation. (2024, January 16). Resist Surveillance Tech, Reject Digi Yatra. Internet Freedom Foundation.

Jaffrelot, C. (2021). Modi's India: Hindu nationalism and the rise of ethnic democracy. Princeton University Press.

Jain, G., & Parsheera, S. (2021). Cinderella's shoe won't fit Soundarya: An audit of facial processing tools on Indian faces. arXiv. https://doi.org/10.48550/arXiv.2112.09326

Jamil, G. (2017). Muslim Women Speak: Of Dreams and Shackles.

Jasanoff, S., & Kim, S.-H. (2009). Containing the atom: Sociotechnical imaginaries and nuclear power in the United States and South Korea. Minerva, 47(2), 119–146. https://doi.org/10.1007/s11024-009-9124-4

Khurana, L, et. al. (2025). *Fintech And Financial Inclusion In India: A Data-Driven Analysis Of Digital Payments And Banking Access.* Journal of Informatics Education and Research.

Kitchin, R. (2014). The data revolution: Big data, open data, data infrastructures and their consequences. Sage.

Krishna Preetham Kanthi. (2024). Privacy, Surveillance, and State Interest: Appraising the DPDP Act through a Constitutional Perspective. Beyond Encryption: Tech & Data Protection, https://www.ijlt.in/post/privacy-surveillance-and-state-interest-appraising-the-dpdp-act-through-a-constitutional-perspect [accessed June 12th, 2024]

Kumar, A. (2022). Ignoring caste and denying development. Data4SDGs.

Marda, V. (2020). Algorithmic accountability in India: A civil society perspective. Medianama.;

Marda, V. (2018). *Artificial Intelligence Policy in India: A Framework for Engaging the Limits of Data-Driven Decision-Making.* Philosophical Transactions A: Mathematical, Physical and Engineering Sciences

NITI Aayog. (2018). *National Strategy for Artificial Intelligence: #AIforAll.* Government of India. Available at: https://www.niti.gov.in

Noble, S. U. (2018). Algorithms of oppression: How search engines reinforce racism. NYU Press.

O'Neil, C. (2016). Weapons of math destruction: How Big Data Increases Inequality and Threatens Democracy. Crown Publishing.

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447-453. https://doi.org/10.1126/science.aax2342

Panigrahi, S. (2022). Marginalized Aadhaar: India's Aadhaar biometric ID and mass surveillance. *ACM Interactions, 29*(2), 16–22.; Frontline. (2024, December 12). Mandatory Aadhaar authentication leads to exclusion of the marginalised from PDS.; The Hindu. (2017, February 18). Aadhaar no standout performer in welfare delivery.

Raji, I. D., & Buolamwini, J. (2019). Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products. Proceedings of the AAAI/ACM Conference on AI Ethics and Society.

Ramachandran Murugesan (2021). Predictive policing in India: Deterring crime or discriminating minorities?. https://blogs.lse.ac.uk/humanrights/2021/04/16/predictive-policing-in-india-deterring-crime-or-discriminating-minorities/ [accessed June 12th, 2024]

Raji, I. D., & Buolamwini, J. (2019). Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products. Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, 429–435. https://doi.org/10.1145/3306618.3314244

Regulation (EU) 2024/1689

Reuters. (2020, November 10). ''Unfair surveillance'? Online exam software sparks global student revolt.' Times of India, reporting on Thomson Reuters Foundation coverage.

Rina Chandran**. (2023).** India's scaling up of AI could reproduce casteist bias, discrimination against women and minorities. https://scroll.in/article/1055846/indias-scaling-up-of-ai-could-reproduce-casteist-bias-discrimination-against-women-and-minorities [accessed June 12th, 2024]

Sambasivan, N., Arnesen, E., Hutchinson, B., Doshi, T., & Prabhakaran, V. (2021). Re-imagining algorithmic fairness in India and beyond. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21) (pp. 315–328). Association for Computing Machinery

Singh, S., & Mohanty, R. (2023). Impacts and ethics of using Artificial Intelligence (AI) by the Indian Police. Police Practice and Research, 24(3), 102–116.

Solove, D. (2008). Understanding privacy. Harvard University Press.

Teltumbde, A. (2018). Republic of caste: Thinking equality in the time of neoliberal Hindutva. Navayana.

Thorat, S., & Neuman, K. (2012). Blocked by caste: Economic discrimination in modern India. Oxford University Press.

UNESCO. (2021). Recommendation on the Ethics of Artificial Intelligence.

United Nations High Commissioner for Human Rights. (2021). The right to privacy in the digital age (A/HRC/48/31). https://www.ohchr.org/en/documents/thematic-reports/ahrc4831-right-privacy-digital-age-report-united-nations-high [accessed June 15th, 2024]

Vaidehi, R., Reddy, A. B., & Banerjee, S. (2021). Explaining caste-based digital divide in India. arXiv.

Winner, L. (1980). Do artifacts have politics? Daedalus, 109(1), 121–136.

Yengde, S. (2019). Caste matters. Viking.

Zuboff, S. (2019). The age of surveillance capitalism. PublicAffairs.

The multiple functions of viral testing during the COVID-19 pandemic in Greece: public health and the governance of society

Katerina Vlantoni[1], Kostas Raptis[1], Athanasios Barlagiannis[2]

[1] National and Kapodistrian University of Athens, Greece

[2] Academy of Athens, Greece

DOI 10.3217/978-3-99161-062-5-009, CC BY 4.0

Abstract. This paper addresses the role of viral testing in the management of the COVID-19 pandemic in Greece by paying attention to the ways testing advanced during the years 2020-21. Prior studies have highlighted the central role of testing in managing the pandemic and its complex implications at both local and global levels. Our analysis focuses on how testing became increasingly widespread and embedded within Greek society over time. Drawing on a range of sources, we situate public health policy-making within the temporal dynamics of the pandemic, tracing the evolving logics and uses of tests, or, as we term it, the distinct functions of testing. Our analysis foregrounds the ways in which tests decoupled from their initial clinical/diagnostic orientation to assume epidemiological, organizational, and punitive functions. In doing so, we highlight the progressive blurring between testing and screening, particularly in relation to the expansion of self-administered rapid antigen tests. We argue that this shift signals not only a transformation in pandemic management but also a broader reconfiguration of health governance toward individual responsibility and self-surveillance. We contextualize these developments within state-led initiatives promoting the digital transformation of public services in Greece. We suggest that the infrastructures and practices surrounding viral testing played a pivotal role in operationalizing this digital agenda. We conclude that the multiple functions of viral testing accumulate in overlapping layers, serving diverse purposes, often simultaneously, not limited to the strict clinical or epidemiological ones.

1 Introduction

From the onset of the COVID-19 pandemic, testing was put forward as a key public health measure. On March 16, 2020, the urge of World Health Organization (WHO) Director-General to 'Test, test, test' made a global impression, affecting the ways international and national public health officials designed and implemented policies to manage the pandemic (WHO, 2020). The complexity in developing testing strategies and interpreting diagnostic tests during health crises, interlinked as they are with social aspects of putting forward widespread testing and/or screening programs, has already been evident in previous cases, such as those involving sexually transmitted infections like syphilis and HIV/AIDS. The COVID-19 pandemic reinstated that testing practices move beyond the laboratory and affect many facets of social and everyday life (Stark, 2020) and brought 'the ethics and politics of medical testing to public attention' (Street and Kelly, 2021, p. 4). In this paper, we draw attention to the multitude of purposes served by viral testing over the course of the COVID-19 pandemic. Through a detailed case study about Greece, where the implementation of widespread testing (i.e. regular asymptomatic screening of large groups of the general population) became a core tenet of the public health policy, we trace the role of testing in the development of public health interventions from the onset of the pandemic (early 2020) until the end of 2021. By focusing on the (public health) logics associated with different uses of viral testing and uses of different tests, we argue that testing assumes multiple functions extending beyond clinical and epidemiological purposes, while these diverse functions often accumulate in overlapping layers.

Studying testing has been a key site of inquiry in history and sociology, as it is crucial in producing diagnoses and informing public health practices, affecting medical practice as well as the notions of health and illness (Jutel, 2009; Armstrong and Eborall, 2012). Tests can be appropriated or used in different ways while diverse tests can serve different functions. Medical testing in an epidemic/pandemic is of crucial importance for clinical purposes (diagnosis and treatment) and for epidemiological purposes (informing public health policies) (for the case of COVID-19, see Beaudevin *et al.*, 2020; Stark, 2020; for the case of HIV/AIDS, see Waldby, 1996, p.105). And, while testing and screening have been considered distinctive functions of tests, we similarly contend that the boundary between the two 'is increasingly becoming blurred' (Petersen and Pienaar, 2021, p. 13).

In the aftermath of COVID-19, a renewed scholarly interest about testing has emerged (see section 2). Already more than thirty years ago, Pinch argued to engage seriously with the sociology of testing, to view 'testing as [a] research site in the sociology of technology' (1993, p. 26). He claimed that '...the sociology of testing should not only be about the subject matter of technology, it should also be about the sorts of social and political relationships embedded within society as a whole' (Pinch, 1993, p. 38). Recently,

Marres and Stark (2020) re-opened the discussion surrounding ‘a new sociology of testing’, significantly broadening the concept itself. They argued that ‘testing in society should be studied from the standpoint of their consequences, that is, on the basis of what tests generate’ (2020, 424). While this use of ‘testing’ with a broader scope can potentially have analytical limitations, it nonetheless prompts us to consider testing as a phenomenon no longer limited solely within the social environment and a specific domain (as in ‘field test’), but one that may involve the ‘very modification of social environments’ (Marres and Stark, 2020, p. 436).

In our effort to identify what tests ‘generate’, we document how COVID-19 public health policy in Greece was shaped by examining the official interventions concerning viral testing. We use the term viral testing (or tests) to capture a range of techniques capable of detecting viral pathogens. Furthermore, we employ it as a broader term that will permit us to extend beyond diagnostic testing (or tests) that serves the main purpose of identifying an infection or disease usually within healthcare settings. In this study, we trace the development of a testing strategy for the management of the pandemic across four phases, from early 2020 to the end of 2021. We focus on the processes through which the use of SARS-CoV-2 tests, particularly as it proliferated with the use of rapid antigen tests (rapid tests, self-tests, among other designations), became widespread beyond the confines of healthcare settings. Through this account of how testing became eventually ubiquitous, we identify four distinct functions of viral testing (of various test types).

In what follows, we begin by presenting our research framework, methods and sources. In the consequent section, we present our account of the public health policy during the COVID-19 pandemic through the lens of testing.

2 Framework/Methodology/Sources

Testing during the COVID-19 pandemic has been a complex issue shaped by a range of factors, many of which have not entirely unprecedented. Infrastructural demands, emergency conditions in overcrowded healthcare facilities, shortages of specialized personnel, disruptions in the global production and distribution of diagnostic consumables, and the need to standardize newly developed (often commercial) tests were among the prevailing challenges. During the dynamic unfolding of the pandemic, research surrounding tests and the development of novel testing technologies was also of outmost importance. Nucleic-acid-based tests (e.g. RT-PCR) were developed from early on to detect the SARS-CoV-2 virus, being considered the ‘gold standard’ (Esbin *et al.*, 2020). PCR-based tests can be labour intensive and time consuming, posing limitations to the scaling up of testing, as it was discussed during 2020. The potential (and pragmatic) use of rapid antigen tests was presented in a comparison like the

following: ‘the best test is not necessarily one that determines whether a person has any evidence of SARS-CoV-2, but one that quickly and accurately identifies individuals who are capable of transmitting the infection to others’ (Manabe *et al.*, 2020). Given the scientific debates surrounding the utility of mass testing with rapid antigen tests, the subtitle of an article in *Nature* (February 2021) read ‘Scientists still debate whether millions of cheap, fast diagnostic kits will help control the pandemic’ (Guglielmi, 2021). Thus, COVID-19 testing strategies varied considerably across countries with respect to clinical and public health uses of different tests (Mina and Andersen, 2021).

Growing scholarly research from the social sciences, including Science and Technology studies, has focused on several aspects surrounding viral testing in the COVID-19 pandemic, commonly through national case studies. For the case of France, an interdisciplinary analysis of the social appropriations of tests in the early phase of the pandemic suggested that ‘the severe limitations of testing infrastructure in France in the first half of 2020 shaped the government’s choice of lockdown strategy’ (Beaudevin *et al.*, 2020, p. 3). In the same vein, Fredriksson and Hallberg (2021) revealed how by targeting specific social groups testing showcased specific organizational and institutional features of Sweden’s National Health System. Fierlbeck *et al*. (2025) demonstrated the complexities in developing a testing strategy foregrounding a multitude of factors, notably non-science ones but institutional, organizational, social and political ones, that became apparent in their comparative analysis of the diversity of COVID-19 testing across four Canadian provinces.

The heightened role of public health uses of viral tests resulted in testing strategies that extended beyond clinical settings, often occurring without the involvement of medical personnel. This is reflected in the dynamic reconfiguration of testing strategies and the incorporation of self-testing in some countries, an issue that we study in this paper for the case of Greece. Petersen and Pienaar (2024) analysed the mass self-testing strategy implemented in Australia during the COVID-19 pandemic, emphasizing the contested role of rapid antigen tests in producing diagnostic certitude while assigning citizens responsibility for self-managing infection risk. Their analysis points to broader implications of self-testing and its subjectification effects, which we do not address here but merit further research. Nonetheless, in the context of the pandemic the decoupling of testing from clinical and epidemiological logics is aligned with a broader reconfiguration of health governance toward individual responsibility and self-surveillance.

Our focus lies on the ways the COVID-19 testing policy developed during the course of the pandemic leading to a mass testing strategy that became diffused within society, comprising of testing in health care facilities and a combination of self-testing and fee-based testing at designated sites. By examining the public health interventions and public policy, we pay attention to the processes of the gradual decoupling of viral testing from its clinical and epidemiological uses following the policy reconfigurations enabling testing to serve a multitude of purposes. In our analysis of the ways testing intersects with the

governance of society during the pandemic, we introduce the concept of 'function' in order to theorize the state-implemented measures from a socio-historical perspective. With this concept, we attempt to capture the purposes and effects that testing can have, meaning the logics associated with the testing policy. In other words, we approach the measures from the standpoint of their potential consequences (see Marres and Stark, 2020), irrespective of the explicit intentions of policymakers or legislators. In this regard, the viral testing functions are analysed as gradually accumulated, with each new layer supplementing rather than displacing the previous ones. We argue that this perspective has merits in order to better understand the role of viral testing during the pandemic governance in Greece, as well as the broader role of testing in public policy and the governance of society.

Fierbleck *et al.* (2025, p. 4) in their recently published article also refer to the functions of testing, explaining that the COVID-19 mitigation measures led to additional functions for testing, which they interchangeably refer to as 'the functions of testing policy'.[1] In their analysis, 'determining which functions testing was to perform' was part of the decision-making process. This bears a difference from our use of the term. We employ testing functions in our analysis to capture both the purposes and the consequences of testing. This permits us to account for the consequences of testing up to the degree of detailing how testing moved beyond the laboratory and healthcare facilities, surpassing clinical and epidemiological functions, and impacting several aspects of social life. Fierbleck *et al.* (2025) share similar findings in terms of the expanding functions of testing in the pandemic governance.

Nonetheless, our approach has limitations. While we are informed by approaches to medical testing claiming that the practices of testing 'have far-reaching socio-political implications, constituting regimes of governance that guide, conduct and shape subjectivities in particular ways and with particular outcomes' (Petersen and Pienaar 2021, p. 8), in this research we do not focus on the experiences of those engaging with testing practices. Along these lines, the effects of testing we refer to are circumscribed by the successive policy interventions, irrespective of instances of contestation, circumvention and tinkering. Regarding the motivation to test, recent research in medical anthropology has advanced our understanding on the ways people engage with tests as 'relational technologies'. For the COVID-19 voluntary asymptomatic testing in Scotland, Bevan *et al*. (2025, p. 289) interviewed participants in such a program and argue that 'testing obligations and responsibilities were experienced as stemming from preexisting relationships to others at multiple scales, rather than being imposed by the state'. Further research in this direction would be illuminating, for instance, by comparing obligatory with

[1] This article was published at the time we were finalizing our manuscript for submission, thus we were not aware of it. We thank Reviewer 2 for pointing it out.

voluntary testing, and possibly further discerning self-testing at home and testing at a public health site.

Our research focuses on the period from the onset of the pandemic in Greece until the end of 2021. We chose this timeframe because it includes the key policy interventions related to viral testing. Further research could extend to the period following the lifting of mandatory testing measures to analyse the public health logics that underpinned those policies. Our analysis draws on mixed sources. On the one hand, our primary material includes publicly available documentary sources. For the period under study, we collected government statements, announcements and press releases related to the deployment of testing public health policy, as well as respective laws and regulations. In addition, we examined news reporting that included interviews and statements of government officials and members of the ad hoc COVID-19 Committee. On the other hand, we draw on observational notes from public events that we attended in the aftermath of the COVID-19 emergency period in which medical professionals, biomedical researchers and public health officials reflected on their experience (2 conferences and 2 individual panels held during 2024-2025, a total of 35 speeches). In addition to the aforementioned sources, we derived insights and information from informal discussions with practitioners who were involved, in various capacities, in different aspects of COVID-19 testing across four institutions. These insights were complemented, to a lesser degree, by our own experiences as Greek citizens engaging with official testing practices during the pandemic.

In the following section, we present our findings. We distinguish four phases in the management of the COVID-19 pandemic in Greece based on the deployment of public health interventions surrounding testing, within the broader context of mitigation measures associated with successive epidemiological waves.

3 Research Findings

3.1 First phase (spring 2020): testing the specific virus

In Greece, the first recorded case of SARS-CoV-2 virus was reported on February 26, 2020. By that time, the emerging epidemic was already a growing concern, with frequent news coverage from China and various European countries. Reporting on the outbreak in neighboring Northern Italy (Gagliano *et al.*, 2020) had a significant impact on public discourse and policymaking processes. Between February 25 and March 30, the Greek government enacted five ‘Acts of Legislative Content’—extraordinary legal instruments issued by the executive under urgent and unforeseeable circumstances—at a pace of nearly one per week. These acts outlined a series of measures to restrict social activities and citizens’ movement in an effort to contain the spread of the virus (see Fig. 1, a

timeline of key events). Educational institutions, cultural venues, and a range of commercial activities were suspended. The Act of March 20 (enacted on March 23) imposed the first nationwide lockdown, initially set for two weeks but extended into early May (Act of Legislative Content, 20 March 2020). Movement outside homes was broadly prohibited, except for six specific reasons and only after notifying authorities by sending an SMS to a designated five-digit number or by carrying a printed certificate to present, if required, at police checkpoints.

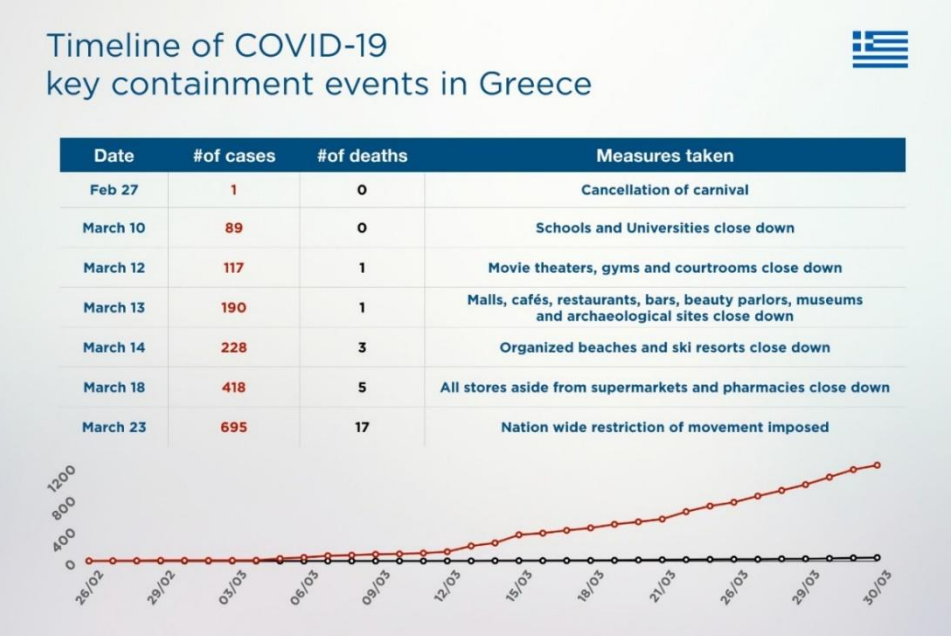


Figure 3 Timeline of key events February – March 2020. (Source: National Public Health Organization)

During this first phase, molecular (PCR-based) testing for SARS-CoV-2 was limited, largely due to infrastructural limitations, shortages of necessary consumables (such as reagents), and an insufficient specialized workforce to support widespread testing. Access to testing was primarily restricted to individuals exhibiting symptoms or those identified through contact tracing efforts. Emergency response units were established by the National Public Health Organization and Civil Protection authorities to implement various containment measures, staffed by specially assigned personnel. Contact tracing was carried out manually by these units, without the aid of digital contact tracing or warning applications. As a result, a key component of identifying ‘possible cases’ of COVID-19 involved samples’ collection conducted at individuals’ homes, targeting symptomatic persons not requiring hospitalization as well as close contacts identified through tracing of confirmed cases. Testing was, of course, also available in hospitals and other healthcare facilities for symptomatic individuals requiring medical care. Additionally, a four-digit hotline was launched to provide round-the-clock information and guidance related to COVID-19.

Access to testing was also contingent on the ability to pay, as private diagnostic centers charged high fees for SARS-CoV-2 tests (Goranitis, 2020).[2] Under strict lockdown measures and severe mobility restrictions, low-wage workers were largely excluded from accessing tests in the private sector. In contrast, those who could afford the cost became the ones *tested* and, consequently, reassured about their infection status. As previously noted, the overall capacity for PCR-based laboratory testing was constrained by the limited availability of essential consumables. In an attempt to meet the end of expanding testing capacity, two significant measures were introduced.

The first was the exceptional launch of a flagship research initiative titled 'Epidemiological study of SARS-CoV-2 in Greece through extensive virus and antibody detection tests, viral genome sequencing and genetic analysis of patients, to address the SARS-CoV-2 virus'.[3] This initiative, assembled on March 28, brought together a consortium of leading research units from four universities and six research centers to conduct, among others, PCR-based testing and viral genome sequencing. One of its key objectives was to meet growing diagnostic demand by developing in-house molecular testing protocols. These protocols were subsequently made available to other laboratories in the broader public sector—ranging from hospitals to research institutions (GSRT, 17.11.2020).

The second measure involved increasing testing capacities through the deployment of automated PCR analyzers, such as those typically used for routine blood screening, at the National Blood Center. In mid-April, the Prime Minister visited the Center and was photographed alongside the newly installed analyzers, stating: 'It is very important that we can add significant testing capacity as we begin to look at gradually reopening society and the economy' (Hellenic Republic, Prime Minister, 15.04.2020).

The lift of the lockdown was decided in early May. By the end of May 2020, the total COVID-19 cases reached 2.917 and the total number of the tests performed concerned 180.518 clinical samples (NPHO, 31.05.2020). As it is obvious, during this phase, tests served mostly clinical purposes and targeted epidemiological surveillance.

3.2 Second phase (June 2020 – January 2021): towards epidemiological screening

Following the gradual removal of restrictive measures in May 2020, the free movement of people inside the country was strongly promoted during the summer. In Greece,

[2] Regarding the national setting, it is important to note the lack of a uniform primary healthcare in Greece. Medical tests and examinations take place, quite extensively, in private diagnostic centers (reimbursed partly by the social security system) and not within the National Health System (see also, Vlantoni, Kandaraki and Pavli, 2017).

[3] For more, see, https://greecevscorona.gr/ (accessed: 10 June 2024).

summer continues to be a significant period for economic activity and the government attempted to implement a framework for both controlling the pandemic and securing the financial activities of the so-called ‘tourist industry’. During this period, testing policies advanced to sporadic screening conducted by mobile units of the National Public Health Organization, which targeted specific high-risk settings, such as social services and elderly care facilities (NPHO, Press release, 04.06.2020).

Special measures were implemented at the country's 'entry points.' Initially, a selective screening process was introduced for passengers arriving at airports, wherein a random sample of incoming passengers was tested. Individuals who tested positive were required to quarantine for two weeks in designated, state-funded hotels (Joint Ministerial Decision, 28.06.2020). In addition, a general entry ban was imposed on foreign nationals from all countries, with the exception of those from EU and Schengen Area member states (Joint Ministerial Decision, 30.06.2020). This policy was framed as a means to control tourist inflow, while effectively keeping borders open to travelers from countries that constitute the core markets for the Greek tourism sector.

Epidemiological surveillance with the use of tests began to increase, as the rapid antigen tests became gradually available. Next to the primary functions of tests to serve clinical and epidemiological surveillance purposes, attempts were carried out to deploy screening. The testing strategy and the public health measures were reassessed following the end of the summer tourist season, as infection rates began to rise. In early November 2020, a series of regional lockdowns was introduced, followed by the implementation of a nationwide lockdown on November 7 (Joint Ministerial Decision, 6.11.2020). Although restrictive, this second lockdown was less stringent than the one imposed during the initial phase of the pandemic.

In the end of 2020, public health policy prioritized expanded testing, particularly through the use of rapid antigen tests (see Fig. 2). This prioritization is also evident in the outcomes of the flagship research initiative (‘Epidemiological study of SARS-CoV-2 in Greece’), which led to the development of the ‘first Greek rapid test for COVID-19’ (Ministry of Development, 30.11.2020). Nonetheless, the domestically developed rapid test was not advanced toward commercial production.

In December 2020, the public health authorities, in collaboration with the Civil Protection and the Ministry of National Defense, launched a platform for epidemiological surveillance titled ‘Form for Free COVID-19 Test’.[4] Through this platform, asymptomatic individuals could register via an online form and express their interest in being tested with antigen rapid tests. Upon registration, individuals would be notified if they were selected and provided with an appointment date at an outpatient facility, typically located within a nearby military hospital. This initiative marked a significant step in shifting public health

[4] Wayback Machine/Web Archive www.testing.gov.gr (date 30/12/2020), accessed: 10 September 2024.

policy toward widespread testing. The stated goal was to conduct epidemiological screening at the community level through 368 designated testing sites nationwide. The government actively promoted participation in the program, framing it as a critical component of the national response to the pandemic. Notably, Prime Minister Kyriakos Mitsotakis publicly endorsed the initiative via a post on the social media platform Twitter, stating: ‘Random sampling to monitor asymptomatic COVID carriers is one aspect of the national strategy against coronavirus. Citizens' assistance in this great effort is of decisive importance. Register here: http://testing.gov.gr’ (Prime Minister GR, 2020).

Simultaneously, mortality rates associated with the pandemic were rising, and concerns were raised regarding the insufficient availability of hospital beds, especially in intensive care units, and the shortage of healthcare personnel. At the end of 2020 and the beginning of 2021, the national vaccination campaign, titled ‘Freedom’, was launched in Greece. The campaign was swiftly endorsed by the media and actively supported by the government, with both the Prime Minister and the President of the Hellenic Republic publicly getting vaccinated (*Kathimerini*, 2020).

Beginning in January 2021, alongside the rollout of mass vaccination, testing efforts were significantly intensified through the expanded use of rapid antigen tests. Mobile units of the National Public Health Organization conducted widespread daily testing across various locations. Notably, this strategy extended beyond traditional healthcare settings, relocating viral testing practices from the confines of the laboratory into public spaces like squares.

Up to this point, we observed a shift from testing symptomatic individuals and their close contacts during the first phase of the pandemic (February to May 2020), to the gradual widespread testing of asymptomatic individuals. Beginning in the summer of 2020, testing was extended to targeted population groups, and by December 2020, it encompassed the general public. In this second phase, testing was increasingly decoupled from the confines of laboratory settings and redeployed across alternative health facilities and public spaces. This spatial reconfiguration of testing was central to the emerging function of epidemiological screening.

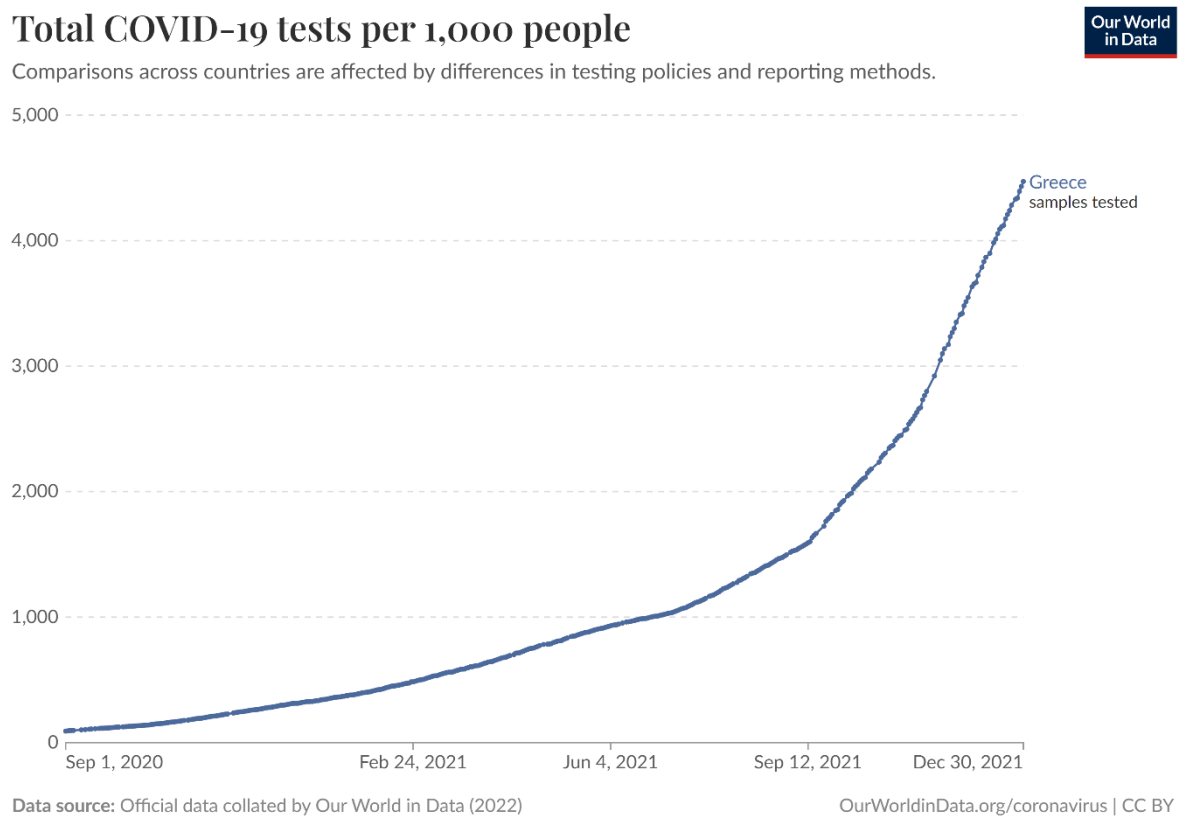


Figure 4 COVID-19 samples tests daily per 1.000 people in Greece, September 1, 2020, to December 31, 2021 (Source: Our World in Data).

3.3 Third phase (February – August 2021): the organizational function of viral testing

The third phase of the pandemic emerged while the mass vaccination campaign was underway and the second lockdown was still in effect. Vaccination was primarily organized by age groups, even if priority was first given to healthcare and political personnel as well as patients with specific underlying health conditions. Eligibility for vaccination gradually expanded, starting with older age groups, allowing individuals to register for an appointment. Vaccination remained voluntary, with the exception of healthcare personnel, for whom it became a work requirement. This process continued over several months, and by the summer of 2021, the vaccine became available to all adults over the age of 18.

As vaccination eligibility broadened, viral testing interventions also expanded. The National Public Health Organization escalated its efforts to expand both PCR-based and rapid antigen testing infrastructures through the establishment of additional mobile units (Joint Ministerial Decision, 2.02.2021). At the same time, a significant move was the free distribution of rapid antigen tests, ‘self-tests’, to every citizen possessing a Social Security Number (Law 4790, 2021). The key argument for this intervention was the re-opening of the society and the economy. On the one hand, the ‘Freedom’ campaign

would gradually lead to a vaccinated and immune to the virus population. On the other hand, the availability of testing would give the opportunity to those awaiting for vaccination a means to act more safely. At this point, testing became a tool for promoting both the vaccination campaign and the broader justification for lifting restrictive measures of the lockdown, meaning reopening the society. In this context, testing served organizational functions for managing in novel ways both the social and economic activities amidst the pandemic crisis. The balance between PCR-based and rapid testing began to shift markedly in favor of the latter (see Table 1).

On March 19, 2021, during the official announcement of the new testing strategy, Akis Skertsos, then Minister of the State, underlined that 'Greece, therefore, based on the new measures that we will introduce, becomes the first country to proceed, from the end of March, to the free provision of individual rapid tests to the entire population of the country. I repeat, free provision of individual rapid tests to the entire population of the country. This way we believe that we will be able to proceed in April with controlled opening of more activities.' (NPHO, Press Release, 19.03.2021).

At the same press conference of March 2021, the Minister of the State referred to broader aspects of the public policy. He specifically referred to the general strategy for the so-called 'digital transformation' of public administration that the Greek state followed from the beginning of the pandemic. In the Act of Legislative Content, enacted on March 23, 2020, which imposed the first lockdown among other measures, the government included the establishment of the 'Single Digital Gateway' and the 'gov.gr' website of the Hellenic Public Administration. 'The state acquires a unified face' was the motto for this new service with its primary goal being the creation of a digital platform designed to gradually integrate a range of essential administrative tasks between the state and its citizens. Kyriakos Pierrakakis, the Minister of Digital Governance since 2019, emphatically declared in May 2020 that SARS-Cov-2 functioned as 'a digital accelerator' for the state policy (*Vouli Watch*, 2020). In March 2021, the Minister of the State elaborated on this strategy as follows: '[…] In the midst of the crisis, we proceeded with the rapid digitalization of the State. From March last year to this year, the digital services provided by the Greek State to citizens have more than doubled, up to 1,138. School registrations, driver's licenses, medical prescriptions, vaccination appointments and much more are now offered digitally, making our lives easier. This is the meaning of digitalization. […]' (NPHO, Press Release, 19.03.2021). Viral testing was also mediated by digital services, especially for the reporting of self-administered test results.

Self-tests were introduced with instructions available in videos in TV, internet and social media. The use of self-tests was promoted as an act of personal and social responsibility; the official message was 'We frequently self-test, we take care of our safety, the health of our loved ones and the lives of our fellow human beings'(Hellenic Government, Press Release, 07/04/2021). Shortly thereafter, undertaking a self-test (still, distributed for free) and reporting the result became mandatory for participation in education activities and

for the workforce in both the public and private sectors, particularly for those required to work on-site (3 Joint Ministerial Decisions of 19.04.2021, Joint Ministerial Decisions of 7.05.2021). To report test results, a new digital platform was set up (self-testing.gov.gr), in which citizens used their tax credentials to identify (as in other services of digital governance). For the minors attending schools, the parents were responsible to report the testing results. For every positive result from a self-test, a confirmatory test was required (PCR or rapid antigen test). Designated sites for confirmatory tests were the testing sites of the National Public Health Organization (PCR or rapid antigen test) and the pharmacies (rapid antigen test).

In May 2021, a digital COVID-19 certificate was introduced for use in a variety of sites, including workplaces, higher education, and certain indoor spaces. This certificate enabled individuals to demonstrate their vaccination status, proof of recent infection, or a recent negative test result as a prerequisite for access. In July 2021, in accordance with EU guidelines aimed at restoring mobility, vaccination certificates were formally issued, and the 'Covid-Free' app was launched (Hellenic Government, 13.7.2021) to verify the health status of travellers (i.e., vaccinated, recently infected, or tested).[5] Within this context, testing functioned as a kind of 'passport', conferring or denying access to specific spaces. The application was used by authorized personnel of entertainment venues, restaurants and cafes, any type of cultural, athletic, festive events (organized either in indoor venues, or in some cases outdoor venues) in order to scan the QR code of the certificate and to permit access to those presenting the digital certificate.

Test results became widely visible and actionable than ever before: for instance, employers in the private sector were granted access to their employees' test statuses and individuals faced penalties if they continued to work even when they had tested positive or skipped regular testing. The same applied for employees in the public sector and for personnel and students in universities (see Joint Ministerial Decisions of 19.04.2021 and of 7.05.2021). During the summer, test results became visible to a range of occasions, throughout each day.

Given the above, testing functioned as an organizational tool for regulating and disciplining key social domains, including educational institutions, workplaces, consumer environments, and the tourism industry. In addition, testing became, on the one hand, an individual obligation requiring self-testing skills, and, on the other hand, increasingly digitalized as test results were shared with and made accessible to a wide array of institutional actors. In this phase, we advocate for a substantial expansion of the functions of tests that permeates both diagnostic testing and screening.

[5] According to Law 4816 (9.07.2021) the Covid-Free application would be compatible with the EU Digital COVID Certificate EUDCC or the equivalent certificate issued from a third country (with a QR code to be scanned for verification).

3.4 Fourth phase (September-December 2021): the punitive function of testing

By the end of summer 2021, the pandemic in Greece entered a new phase marked by the emergence and rapid spread of a new variant of SARS-CoV-2, the Delta variant. This development reignited public discourse on the appropriate public health measures to be implemented. At the same time, the vaccination campaign had reached a point where all adults were eligible to register and schedule appointments for vaccination. Public opposition began to manifest more visibly, with demonstrations and critical debates surrounding the perceived mandatory, either explicitly enforced or implicitly pressured, nature of COVID-19 vaccination policy (*Kathimerini*, 2021). In response, the government introduced targeted measures aimed at different social and age groups. For younger individuals, a policy known as the 'Freedom Pass/Data' was enacted, offering 50GB of free mobile data to those turning 18 in that year, on the condition that they had been vaccinated (Joint Ministerial Decision, 9.11.2021). For older citizens, specifically those over the age of 60, the government introduced a fine of 100 euros per month if they chose to remain unvaccinated (Law 4865, 2021).[6]

These age-specific policies reflected an effort to incentivize vaccination through both reward and penalty, signaling a shift in the state's approach from encouraging voluntary participation to enforcing compliance. By September 2021, government's discourse increasingly depicted unvaccinated individuals as having wilfully refused vaccination, given that access to COVID-19 vaccines had become broadly available to the entire adult population. This framing positioned vaccine hesitancy not as a matter of limited access or uncertainty, but as a deliberate act of non-compliance with public health imperatives. At that time, the Greek government introduced a policy mandating regular testing for COVID-19 at the expense of unvaccinated employees (NPHO, Press Release, 24.08.2021). According to the new measures, formalised in Joint Ministerial Decisions, as of September 13 those employed in the private sector and physically present at their workplace were required to undergo a test (either PCR or rapid antigen test) once per week (or twice in some cases) bearing the relevant cost (Joint Ministerial Decision, 16.10.2021). The cost of testing was set at approximately 10 euros per rapid antigen test and could be carried out at private diagnostic laboratories, clinics and pharmacies.[7] This measure resulted in an estimated monthly cost of approximately 40 euros for unvaccinated workers, effectively adding an economic burden that functioned as an indirect form of pressure to comply with vaccination requirements.

[6] The regulation on fines remained in effect until 2022, while two years later, in 2024, the Ministry of Health waived the fines for those who had not paid them by then (*Kathimerini,* 2024).

[7] The cost of testing varied depending on the testing site and the type of test. The great majority of individuals would choose rapid tests to be carried out at a pharmacy that was the less costly option.

In response to widely publicized cases in which unvaccinated citizens circumvented costly mandatory testing by paying for fraudulent proof of recent infection, the type of accepted test was further specified. In early December 2021, the government mandated that the declaration of SARS-CoV-2 infection could only be made following a positive laboratory-based PCR test (Ministry of Health, 2021). Unvaccinated citizens were required to undergo testing in a private diagnostic center (unless they presented symptoms and were in need of hospitalization), at a cost approximately six times higher than that of a rapid antigen test.

Still, demonstration of proof of vaccination, of recent infection or of a negative test result was required in retail settings, healthcare facilities and social, cultural activities to permit entry indoors. For those vaccinated, access to testing (free of charge) was granted if they had symptoms or when they presented themselves voluntarily at the testing sites of the National Public Health Organization. We should note that at that point self-tests were also being sold in pharmacies. Thus, testing at home (getting tested in order to declare the result or on one's free will) gradually became a common practice (see, Table 1). In light of the Christmas festivities and the concerns arising from the emergence of the SARS-CoV-2 Omicron variant, the Ministry of Health distributed a free rapid antigen test to every adult citizen, regardless of vaccination status, during the week of 6–11 December 2021, for voluntary use toward epidemiological monitoring (Ministry of Health, 2021).

Samples Tested for SARS-CoV-2 (per mode of testing)			
Testing period	Laboratory tests (RT-PCR)	Rapid Ag (by NPHO in designated sites)	Declared self-tests (Rapid Ag)
1/1/2020 - **31/12/2020**	2.803.026	579.462	-
1/1/2020 - **31/03/2021**	4.171.213	2.364.533	-
1/1/2020 - **30/09/2021**	6.632.532	13.506.241	38.972.750
1/1/2020 - **31/12/2021**	8.282.716	38.966.229	66.949.593

Table 1.: The table presents the samples tested, according to the designated Daily reports of the National Public Health Organization (NPHO, Daily report, 31/12/2020, 31/03/2021, 30/09/2021 and 31/12/2021).

During this phase, the functions of testing multiplied serving additional purposes. The requirement for regular testing of unvaccinated workers was not merely a way to incentivize vaccination but it assumed a punitive function for those who had been labeled 'unvaccinated'. It targeted specific groups—primarily low-waged workers—by imposing

them a further financial burden: individuals were compelled either to comply with vaccination mandates or to bear the recurring cost of mandatory testing, thereby quite literally paying for their choices.

In December 2021, the government launched the digital application 'Gov.gr Wallet,' which enabled users to store and present COVID-19 vaccination and testing certificates (*TA NEA*, 2021). The innovative aspect of this initiative was the rapid integration of the national identity card into the same application, establishing it as an official tool for digital identification. Over time, the application was progressively expanded to include additional state-issued documents, such as driver's licenses. At present, the Gov.gr Wallet serves as an official platform hosting a broad array of personal identification documents. Notably, one of its more recent applications includes the purchase of football match tickets, a function introduced in response to new legislation aimed at strengthening personal identification and enhancing security protocols at sporting events.

4 Discussion – Layers of testing

This paper showcased the functions of viral testing within the public health policies implemented during the COVID-19 pandemic in Greece. We introduced the concept of functions to interpret the testing interventions from the standpoint of their potential consequences in society as a whole. The four distinct functions of various types of tests accumulated progressively over time in overlapping layers, with each new layer supplementing rather than displacing the previous ones. Our aim was to reveal the processes through which testing has become a ubiquitous feature of everyday life. Further research is needed for assessing the value of this approach for different societal groups.[8]

The first function appeared at the outset of the pandemic (and is in place to this day). It was oriented toward clinical diagnosis and targeted individuals presenting symptoms, with the aim of confirming infection and guiding clinical intervention. We refer to this function as *testing the specific virus*, a function rooted in biomedical logic and healthcare provision.

As the virus spread, a second function gradually took shape. Initially introduced sporadically during the summer of 2020, it became more institutionalized by December

[8] *It is important to note that during the pandemic, there were always people that were targeted or excluded, directly or indirectly, by the public policies; in this paper our research does not expand to cover this issue. For marginalized populations—such as undocumented migrants, whether residing in camps or in urban settings, and homeless people—access to testing and screening policies varied. Many among these groups lacked a Social Security Number or access to the digital platform gov.gr, both of which were prerequisites for participation in testing. Further research is needed that can focus on the potentially discriminatory character of the emergency measures (for instance, targeted screening programs).*

of that year. This function focused on *testing for epidemiological screening*, expanding from group-based epidemiological surveillance to population-level screening, shifting the focus away from individual diagnosis.

With the launch of the mass vaccination campaign, viral testing acquired additional functions, signaling further purposes as Pinch (1993) might have pointed out. Gradually, a blurring occurred between diagnostic testing and epidemiological screening. This blurring, we argue, indicates that testing functioned in new social areas as an *organizational tool*, beyond its initial biomedical and/or epidemiological logics. The mass use of self-tests was promoted both as a way to self-diagnose and self-manage one's health, and as an invaluable contribution to epidemiological screening. In this expanded capacity, testing also functioned as an apparatus of coordination and control within workplaces, educational institutions, hospitals, and other public settings. Within the context of the government's 'Freedom' campaign for COVID-19 vaccination, testing became part of a broader disciplinary mechanism aimed at regulating mobility, individual behavior, civic responsibility and social relations. In this sense, it served as an infrastructural intermediary that helped sustain institutional operations and social relations under pandemic conditions by reordering them.

However, contestation quickly emerged. During the fourth phase, we argue that testing assumed a *punitive* function, particularly in relation to unvaccinated individuals. No longer serving primarily clinical or epidemiological purposes, mandatory testing was repurposed as a tool of sanction. As such, it functioned not to persuade or protect, but to punish, both symbolically and materially, those who refused vaccination.

At this point, it is important to foreground an underlying and persistent dimension that remained present throughout the entire period under study. This is what the government called 'digital transformation of public administration,' a policy objective that had already been decided before the pandemic emerged. The pandemic functioned as an 'opportunity' for the Greek government to pursue this transition, while the emergency measures (lockdown, testing, vaccination) were also implemented through this infrastructural change. Computing integration into public infrastructures signified a shift in health policy or, as Agar (2003) probably would argue, a re-appearance of the state in social life in spaces where previously it was absent.

Considering the above, the dynamic configuration of the mass testing strategy, exemplified in the widespread use of self-tests and rapid antigen tests at designated sites, enabled the embedding of testing within public spaces, domestic settings, and everyday life. Beyond the blurring of boundaries between diagnostic testing and epidemiological screening, the organizational and punitive functions rendered widespread testing a prerequisite for governance. Further research could explore whether such processes of diffusing testing into society encompass an educational function, that of cultivating a culture of testing. Self-tests form part of a wider shift toward

the self-management of health and the reinforcement of individual responsibility, aligned with deeper political objectives. The pandemic provided an opportunity to observe the shifting priorities favoring private profit over collective welfare, as reflected in the government's reluctance to provide substantial support for healthcare personnel and to invest in the public health system's infrastructures.

Viral testing as implemented within public health policies during the pandemic in Greece exhibited flexibility, accommodating diverse purposes, objectives, and strategies rather than functioning solely as clinical or epidemiological intervention. Our analysis of the four overlapping functions of viral testing, from diagnostic and epidemiological to organizational and punitive, demonstrates how it simultaneously advanced strategies decided long ago, such as digital integration, and responded to emerging challenges, including contestation and vaccine refusal. By taking testing as our unit of analysis, we argue for its significance as a critical site for investigating broader social processes and the governance of everyday life in contemporary society.

Acknowledgments

The authors would like to express their gratitude to the two anonymous reviewers for their thoughtful comments and suggestions for improvement. This research was undertaken in the context of the research project 'Testing under crisis, a history from HIV/AIDS to COVID-19: between public debates and health policies – CrisisTesting', carried out within the framework of the National Recovery and Resilience Plan Greece 2.0, funded by the European Union NextGenerationEU (Implementation body: Hellenic Foundation for Research and Innovation).

References

‘Act of Legislative Content of 20 March 2020 on the urgent measures to address the consequences of the risk of spreading the coronavirus COVID-19, to support society and entrepreneurship, and to ensure the smooth functioning of the market and public administration’ (2020) *Government Gazette* A’ 68.

Agar, J. (2003) *The Government Machine: A Revolutionary History of the Computer.* Cambridge, Massachusetts: The MIT Press.

Armstrong, N. and Eborall, H. (2012) ‘The sociology of medical screening: past, present and future’, *Sociology of health & illness*, 34(2), pp 161–176. Available at: https://doi.org/10.1111/j.1467-9566.2011.01441.x.

Beaudevin, C. *et al.* (2021) ‘‘Test, Test, Test!’: Scarcity, Tinkering, and Testing Policy Early in the COVID-19 Epidemic in France’, *Medicine Anthropology Theory*, 8(2), pp. 1–31. Available at: https://doi.org/10.17157/mat.8.2.5116.

Bevan, I., Bauld, L., and Street, A. (2024) ‘Who We Test For: Aligning Relational and Public Health Responsibilities in COVID-19 Testing in Scotland’, *Medical Anthropology*, 43(4), pp. 277–294. Available at: https://doi.org/10.1080/01459740.2024.2349514.

Esbin, M. N. *et al.* (2020) ‘Overcoming the bottleneck to widespread testing: a rapid review of nucleic acid testing approaches for COVID-19 detection’, *RNA*, 26(7), pp. 771–783. Available at: https://doi.org/10.1261/rna.076232.120.

Fierbeck, K. *el al.* (2025) ‘Testing ‘the science’: A comparative analysis of COVID-19 testing policy across four Canadian provinces’, *Social Science and Medicine*, 371, 117880. Available at: https://doi.org/10.1016/j.socscimed.2025.117880.

Fredriksson M. and Hallberg A. (2021) ‘COVID-19 Testing in Sweden During 2020-Split Responsibilities and Multi-Level Challenges’, *Front Public Health*, 19 (9), 754861.

Gagliano, A. *et al.* (2020) ‘COVID-19 Epidemic in the Middle Province of Northern Italy: Impact, Logistics, and Strategy in the First Line Hospital’, *Disaster medicine and public health preparedness*, 14(3), pp. 372–376. Available at: https://doi.org/10.1017/dmp.2020.51.

General Secretariat for Research & Technology (GSRT), Ministry of Development & Investment, Document N. 121933 - 17-11-2020 ‘Ερώτηση με αριθμό πρωτ. 1017/26-10-2020’ about ‘Diagnostic tests for the coronavirus pandemic’, signed by the General Secretary for Research & Technology.

Goranitis, G. (2020) ‘All you want to know about coronavirus tests’ (Όλα όσα θέλατε να μάθετε για τα τεστ του κορονοϊού). *Inside Story*, 19 April. Available at: https://insidestory.gr/article/covid19-test-koronoios.

Guglielmi G. (2021) 'Rapid coronavirus tests: a guide for the perplexed', *Nature*, 590(11), pp. 202-205.

Hellenic Government, Press Release (2021). *Self test – simple and easy. User's guide*, 7 April 2021. [Online]. Available at: https://www.government.gov.gr/self-test-efkola-ke-apla-odigies-chrisis/ (Accessed 15 June 2025).

Hellenic Government (2021). *The 'Covid Free GR' Application is Available*. Public Announcement, 13 July [Online]. Available at: https://www.youtube.com/watch?v=7CLJPf0u3KY&ab_channel=%CE%95%CE%BB%CE%BB%CE%B7%CE%BD%CE%B9%CE%BA%CE%AE%CE%9A%CF%85%CE%B2%CE%AD%CF%81%CE%BD%CE%B7%CF%83%CE%B7 (Accessed: 15 June 2025).

Hellenic Republic, Prime Minister (Press Release). 'Prime Minister's visit to the National Blood Center' (Επίσκεψη του Πρωθυπουργού Κυριάκου Μητσοτάκη στο Εθνικό Κέντρο Αιμοδοσίας). *Prime Minister Official Website*, 15 April 2020. Available at: https://www.primeminister.gr/2020/04/15/23766.

'Joint Ministerial Decision of the Ministers of Finance – Development and Investments – Civil Protection – Labour and Social Affairs – Health – Interior – Infrastructure and Transport – Shipping and Island Policy No. D1a/GP.oik. 40383 of 28 June 2020 on the imposition of the measure of sample laboratory testing and temporary restriction of persons entering from abroad, in order to limit the spread of coronavirus COVID-19' (2020) *Government Gazette* B' 2602.

'Joint Ministerial Decision of the Ministers of Civil Protection – Health – Interior No. D1a/GP.oik. 41013 of 30 June 2020 on the imposition of the measure of entry ban into the country for nationals of third countries, excluding those of the European Union and the Schengen Agreement, in order to limit the spread of coronavirus COVID-19, for the period from 1.7.2020 to 15.7.2020' (2020) *Government Gazette* B' 2658.

'Joint Ministerial Decision of the Ministers of Finance – Development and Investments – Civil Protection – National Defense – Education and Religious Affairs – Labour and Social Affairs – Health – Environment and Energy – Culture and Sports – Justice – Interior – Migration and Asylum – Infrastructure and Transport – Shipping and Island Policy – Rural Development and Food No. D1a/GP.oik. 71342 of 6 November 2020 on the emergency measures for the protection of public health from the risk of further spread of coronavirus COVID-19 throughout the national territory, for the period from Saturday, November 7, 2020, until Monday, November 30, 2020' (2020) *Government Gazette* B' 4899.

‘Joint Ministerial Decision of the Ministers of Development and Investments – Health – Interior No. D1a/GP.oik. 3055 of 2 February 2021 on the establishment of Mobile Health Units for Special Molecular Testing Purposes for the immediate execution of SARS-CoV-2 rapid antigen tests to detect COVID-19 cases (Special Purpose Mobile Molecular Testing Units – K.O.M.Y.)’ (2021) *Government Gazette* B’ 387.

‘Joint Ministerial Decision of the Ministers of National Defense – Health – Justice – Interior – State No. D1a/GP.oik. 24527 of 19 April 2021 on the implementation of the mandatory measure of diagnostic testing for COVID-19 infection for judicial and prosecutorial officers and military judges’ (2021) *Government Gazette* B’ 1582.

‘Joint Ministerial Decision of the Ministers of Education and Religious Affairs – Labour and Social Affairs – Health – Justice – Interior – Digital Governance – State No. D1a/GP.oik. 24526 of 19 April 2021 on the implementation of the mandatory measure of diagnostic testing for COVID-19 infection for public sector employees providing work in person at their workplace’ (2021) *Government Gazette* B’ 1583.

‘Joint Ministerial Decision of the Ministers of Finance – Development and Investments – Labour and Social Affairs – Health – Infrastructure and Transport – Shipping and Island Policy – State No. D1a/GP.oik. 24525 of 19 April 2021 on the implementation of the mandatory measure of diagnostic testing for COVID-19 infection for private sector employees providing in-person work’ (2021) *Government Gazette* B’ 1588.

‘Joint Ministerial Decision of the Ministers of Development and Investments – Education and Religious Affairs – Labour and Social Affairs – Health – Interior – State No. D1a/GP.oik. 28259 of 7/7 May 2021 on the implementation of the mandatory diagnostic testing measure for COVID-19 infection for students, academic, and other personnel of higher education institutions (2021) *Government Gazette* B’ 1866.

‘Joint Ministerial Decision of the Ministers of Finance – Health – Digital Governance – State No. 4700 of 9 October 2021 on the procedures for granting the *Freedom Pass/Data*’ (2021) *Government Gazette* B’ 4675.

‘Joint Ministerial Decision of the Ministers of Finance – Development and Investments – Education and Religious Affairs – Labour and Social Affairs – Health – Culture and Sports – Justice – Interior – Digital Governance – Infrastructure and Transport – Shipping and Island Policy – Tourism – State – Deputy Minister to the Prime Minister No. D1a/GP.oik. 64232 of 16 October 2021 on the implementation of the mandatory measure of diagnostic testing for COVID-19 infection for private sector employees providing in-person work at the workplace’ (2021) *Government Gazette* B’ 4766.

Jutel, A. (2009) ‘Sociology of diagnosis: a preliminary review’, *Sociology of Health & Illness*, 31, pp. 278–299. Available at: https://doi.org/10.1111/j.1467-9566.2008.01152.x.

Kathimerini (2020). *The first vaccinations against Covid-19 in Greece are a fact*, 27 December 2020. Available at: https://web.archive.org/web/20210304100716/https://www.kathimerini.gr/society/561208456/gegonos-o-protos-emvoliasmos-stin-ellada-kata-tis-covid-19/ (Accessed 15 June 2025)

Kathimerini (2021) ‘Protests Against Vaccines as the Delta Variant Gallops Ahead’, Kathimerini, 14 July. Available at: https://www.kathimerini.gr/society/561433585/sygkentroseis-kata-ton-emvolion-me-tin-metallaxi-delta-na-kalpazei/ (Accessed: 15 June 2025).

Kathimerini (2024) ‘Georgiadis for coronavirus: Fines for unvaccinated elderly people are being written off’, 10 January 2024. Available at: https://www.kathimerini.gr/life/health/562821970/georgiadis-gia-koronoio-diagrafontai-ta-prostima-stoys-anemvoliastoys-ilikiomenoys/ (Accessed 15 June 2025).

‘Law 4683 of 10 April 2020 on the ratification of the Presidential Act of 20 Mars 2020 ‘Urgent measures to address the consequences of the risk of spreading the coronavirus COVID-19, to support society and entrepreneurship, and to ensure the smooth functioning of the market and public administration’ (Government Gazette A’ 68) and other provisions’ (2020) *Government Gazette* A’ 83.

‘Law 4790 of 31 Mars 2021 on the urgent provisions for the protection of public health from the ongoing consequences of the COVID-19 coronavirus pandemic, development, social protection, and the reopening of courts and other matters’ (2021) *Government Gazette* A’ 48.

‘Law 4865 of 4 December 2021 on the establishment and organization of a legal entity under private law under the name 'National Central Health Procurement Authority', on the strategy for central procurement of health products and services, and on other urgent provisions for public health and social welfare’ (2021) *Government Gazette* A’ 238.

Manabe Y.C., Sharfstein J.S. and Armstrong K. (2020) ‘The Need for More and Better Testing for COVID-19’, *JAMA*, 324(21), pp. 2153–2154.

Marres, N. and Stark, D. (2020) ‘Put to the test: For a new sociology of testing’, *The British journal of sociology*, 71(3), pp. 423–443. Available at: https://doi.org/10.1111/1468-4446.12746.

Mina, J. M, and Andersen, G. K. (2021) 'COVID-19 testing: One size does not fit all', *Science*, 371, pp. 126-127. Available at: https://www.science.org/doi/10.1126/science.abe9187.

Ministry of Development (2020) 'The First Greek Rapid Test for COVID-19 is a Reality', Press Release, 30 November. Available at: https://www.mindev.gov.gr/36151/ (Accessed: 15 June 2025).

Ministry of Health (2021) 'Announcements from the Minister of Health Thanos Plevris, the President of EODY Theoklis Zaoutis, and the General Secretary of PHC Marios Themistocleous', Press Release, 2 December. Available at: https://www.moh.gov.gr/articles/ministry/grafeio-typoy/press-releases/9829-anakoinwseis-apo-ton-ypoyrgo-ygeias-thano-pleyrh-ton-proedro-toy-eody-theoklh-zaoyth-kai-ton-geniko-grammatea-pfy-mario-themistokleoys (Accessed: 15 June 2025)

National Public Health Organization (NPHO). *Daily epidemiological surveillance report of infection by the novel coronavirus (COVID-19)*. 31 May 2020.

National Public Health Organization (NPHO). *Daily epidemiological surveillance report of infection by the novel coronavirus (COVID-19)*. 31 December 2020.

National Public Health Organization (NPHO). *Daily epidemiological surveillance report of infection by the novel coronavirus (COVID-19)*. 30 September 2021.

National Public Health Organization (NPHO). *Daily epidemiological surveillance report of infection by the novel coronavirus (COVID-19)*. 31 December 2021.

National Public Health Organization (NPHO), Press Release (2020). *The Special Purpose Mobile Molecular Testing Units – K.O.M.Y. of the National Public Health Organization exceeded 1,000 dispatches during their first month of operation*, 4 June 2020. [Online]. Available at: https://eody.gov.gr/oi-komy-toy-eody-xepernoyn-tis-1-000-apostoles-kata-ton-proto-mina-leitoyrgias-toys/ (Accessed: 15 June 2025).

National Public Health Organization (NPHO), Press Release (2021). *Briefing of accredited journalists by the Deputy Minister for Civil Protection and Crisis Management Nikos Hardalias, Professors Vana Papaevangelou and Gikas Majorkinis, and the Deputy Minister to the Prime Minister Akis Skertsos*, 19 March 2021. [Online]. Available at: https://eody.gov.gr/enimerosi-20210319/.

National Public Health Organization (NPHO), Press Release (2021). *Announcements by the Minister of Health Vasilis Kikilias regarding public health measures concerning unvaccinated citizens*, 24 August12021. [Online]. Available at: https://eody.gov.gr/anakoinoseis-ypoyrgoy-ygeias-vasili-kikilia-gia-ta-metra-dimosias-ygeias-poy-aforoyn-se-anemvoliastoys-polites/.

Petersen, A. and Pienaar, K. (2021) ‘Testing for Life? Regimes of Governance in Diagnosis and Screening’, *Science, Technology and Society*, 26(1), pp. 7-23. Available at: https://doi.org/10.1177/0971721820964889.

Petersen, A. and Pienaar, K. (2024) ‘Competing realities, uncertain diagnoses of infectious disease: Mass self-testing for COVID-19 and liminal bio-citizenship’, *Sociology of Health & Illness*, 46(S1), pp. 242–260. Available at: https://doi.org/10.1111/1467-9566.13694.

Pinch, T. (1993) ‘‘Testing - One, Two, Three... Testing!’: Toward a Sociology of Testing’, *Science, Technology, & Human Values*, 18(1), pp. 25–41.

PrimeministerGR (2020) Post, 22 December, Twitter (now X). Available at: https://x.com/PrimeministerGR/status/1341379506175676417 (Accessed: 10 April 2023).

Stark, D. (2020) ‘Testing and Being Tested in Pandemic Times’, *Sociologica*, 14(1), pp. 67–94. Available at: https://doi.org/10.6092/issn.1971-8853/10931.

Street, A. and Kelly, A. H. (2021) 'Introduction: Diagnostics, medical testing, and value in medical anthropology', *Medicine Anthropology Theory*, 8(2), pp. 1-16. Available at: https://doi.org/10.17157/mat.8.2.6516.

TA NEA (2021) ‘Covid Free Wallet – Identity and Certificate in One – See the Activation Instructions’, 28 December. Available at: https://www.tanea.gr/2021/12/28/science-technology/covid-free-wallet-taytotita-kai-pistopoiitiko-se-ena-deite-tis-odigies-energopoiisis/ (Accessed: 14 June 2025).

Vlantoni, K., Kandaraki, A., and Pavli, A. (2017) ‘Medical Technologies and Health Policies in Post WWII Greece’, *History of Technology (Special Issue: History of Technology in Greece, from the Early 19th to 21st Century)*, 33, pp. 107-133.

Vouli Watch (2020) ‘K. Pierakakis: A Digital Accelerator for COVID-19’, Vouli Watch, 26 May. Available at: https://vouliwatch.gr/news/article/k-pierakakis-psifiakos-epitahyntis-covid-19 (Accessed: 15 June 2025).

Waldby, C. (1996) *AIDS and the Body Politic: Biomedicine and Sexual Difference*. Routledge.

WHO (2020) *WHO Director-General’s opening remarks at the media briefing on COVID-19* - 16 March 2020. Available at: https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---16-march-2020 (Accessed 10 June 2025).

Research Ethics Governance with Responsible AI Sandboxes

Michael Gille, Marina Tropmann-Frick

Hamburg University of Applied Sciences, Germany

DOI 10.3217/978-3-99161-062-5-010, CC BY 4.0

Abstract. University research ethics committees (REC) face challenges in overseeing artificial intelligence (AI) research. Historically rooted in biomedical and social science paradigms, REC were not designed to evaluate the epistemic, temporal, and normative complexities of AI and machine learning research. The EU's AI Act exacerbates this tension by exempting academic research from its scope while at the same time promoting the application of ethics guidelines, thereby creating a zone of normative ambiguity. This paper critically examines the resulting governance vacuum. We argue that conventional ethical review processes are inadequate in many cases for reasons inherent in AI research, which is often iterative and interdisciplinary, characterized by shifting goals and emerging risks, as well as because of the normative and socio-technical co-construction of AI technology development. We propose the *Responsible Artificial Intelligence Sandbox* as a model for research ethics governance. It reframes the role of REC from static evaluators to co-constructors of ethical oversight within experimental research environments. Drawing on insights from regulatory sandboxes in EU law and national contexts, this conceptual model enables dynamic, participatory, and reflexive engagement with ethics throughout the research lifecycle. Two main contributions are made: we diagnose structural misalignments of existing research ethics infrastructure and conceptualize responsible AI sandboxes as an institutional and methodological innovation that aligns ethical governance with the nature of research on and with AI.

1 Introduction

University-based Research Ethics Committees (REC), Institutional Review Boards (IRB), and Ethics Review Committees (ERC), in the following collectively referred to as REC, are set up to ascertain ethically responsible research oversight. Rooted historically in biomedical, behavioural, and social research paradigms, these bodies were institutionalized to protect human subjects and uphold normative standards of scientific integrity (Shamoo and Resnik, 2009). However, the rapid emergence of data processing, algorithmic, machine/deep learning (in the narrower sense) and artificial intelligence (AI) research (in the wider sense), hereafter jointly referred to as AI research, have exposed

profound limitations in established research ethics governance models (Stahl et al., 2025; Hadley et al., 2025; Bouhouita-Guermech et al., 2023; Petermann et al., 2022; González-Esteban and Calvo, 2021; Ferretti et al., 2021). As computational systems become both the objects and instruments of inquiry, REC increasingly face tasks they have initially neither conceptually nor procedurally been designed to address. At the same time, the European legislator takes a cautious approach to AI research and grants university research far-reaching freedom. The EU's Artificial Intelligence Act (AI Act), introducing a risk-based regulatory framework for AI development and deployment, explicitly exempts academic AI research from its direct applicability (AI Act, Art. 2(6), rec. 25). This exemption produces a structurally intended zone of governance ambiguity. Mandated to maintain the high standards of biomedical ethics reviews (Brenneis et al., 2024), universities thus find themselves in a paradoxical position: they are encouraged to advance high-risk AI research under conditions of ethical autonomy, while lacking institutional mechanisms and resources tailored to the novel and recursive dilemmas this research entails.

AI research often involves open-ended exploration, emergent objectives, and shifting definitions of key principles such as fairness, trustworthiness, explainability, and human-centeredness (Díaz-Rodríguez et al., 2023). In such volatile research settings, ethical and legal oversight cannot be a static, one-time assessment. Yet traditional review processes struggle to keep pace, often reduced to bureaucratic hurdles by technical researchers, or overwhelmed by the sheer complexity of innovation (Brenneis and Burden, 2024). The governance landscape is increasingly saturated with normative frameworks, ranging from 'trustworthy AI' to data protection principles and fundamental rights mandates (Jobim et al, 2019), but these frameworks often operate in silos, lack enforceability, or offer only abstract guidance (Resseguier, 2024; González-Esteban and Calvo, 2024). What is absent is a tailored mode of ethics governance that is dynamic, participatory, and capable of engaging with uncertainty that research 'on' and 'with' AI brings about. This paper proposes the concept of the *Responsible AI Sandbox* as a model of integrative research ethics governance. Sandboxes, as spaces for regulatory and technical experimentation, offer an architectural shift: they enable REC not only to assess, but to enable the *co-construction* of ethical and governance practices within the experimental settings of AI innovation itself, while at the same time reducing the risk of evaluation gaps.

This paper makes two contributions to ongoing debates in science and technology studies (STS), research governance, and AI research ethics, providing specific impetus for a normative reflection and governance of AI research (research that develops and/or uses AI): *First,* it critically diagnoses the dilemmatic structural inadequacies of traditional ethics-approval mechanisms when confronted with the epistemic, temporal, and normative complexities of AI research. In this regard, special attention is paid to the co-constructive character of responsible AI research, in which the very definitions of risk,

fairness, trustworthiness, and accountability are shaped within the research process itself, rather than being fully specifiable ex ante. By stressing this entanglement of ethics and epistemics, the paper contributes to STS and RRI scholarship on the co-production of algorithmic knowledge and normativity, while offering a fresh institutional diagnosis of selected pressures REC currently face. *Second,* the paper introduces and conceptualizes the *Responsible Artificial Intelligence Sandbox* (RAIS) as an integrative governance model that creates not just a space to engage with ethical norms, but enables dynamic, participatory, and reflexive forms of oversight within university AI research settings. At the same time, this sandbox model is positioned as a response to the governance vacuum created by the EU AI Act's exemption of research activities, and as a concrete mechanism for enabling regulatory learning through an open and normatively reflective research practice within academic institutions. Framed as both a methodological and institutional innovation, this sandboxing approach allows universities to experiment with ethical and legal frameworks together with technological development, thus transforming them into laboratories of governance modalities themselves.

This paper is guided by a set of research questions articulating a broader inquiry into how universities can adapt their governance infrastructures to align with the epistemic and normative demands of responsible AI research, while remaining grounded in principles of academic freedom and anticipatory accountability. The following questions structure the inquiry:

- What *structural* limitations and *epistemic* mismatches do REC face when tasked with the oversight of AI research?
- How should a 'Responsible AI Sandbox' be conceptually designed to meet key ethical and regulatory imperatives such as responsibility, fairness, trustworthiness, and risk mitigation and what insights can, in this regard, be drawn from regulatory sandbox approaches in EU law and national jurisdictions through analogical reasoning and governance transfer?

This paper employs a conceptual research methodology grounded in interpretive STS and regulatory studies, synthesising insights from policy documents, scholarly literature including comparative case analyses of regulatory sandboxes. It uses abductive reasoning to explore how institutional design elements can be transposed into intra-university governance frameworks for responsible AI. The approach highlights normative and epistemic dimensions of experimentation, focusing on co-construction, reflexivity, and anticipatory governance. Through critical analysis, analogical reasoning and institutional comparison, the paper seeks to translate and adapt principles of experimental regulation to the university context, with the aim of developing a conceptual framework for responsible AI sandboxes.

This paper is organized as follows. Section 2 identifies relevant strands of literature on the ethical governance of AI research. This is followed by methodological considerations (section 3). The subsequent section 4 lays out selected structural pitfalls in the ethical governance of AI research before the paper then turns to the responsible AI sandbox (section 5). We conclude with an outlook in section 6.

2 Related Work

This paper´s point of departure are approaches and reflections on AI research ethics governance. Owing partly to AI research leading to critical scrutiny and criticism of the 'traditional' research ethics governance set-up in universities, there have been calls for reforms (Masso et al., 2025; Gonzáles-Esteban and Calvo, 2022; Petermann et al., 2022). Increasing attention is being paid to the shifting role of REC in AI-related research, regarding both research on and with AI (Esmaili et al., 2025; Brenneis/Burden, 2024; ZEVEDI, 2023). For purposes of this paper´s integrative approach, two related strands of discussion stand out and are put into focus: There are calls for an interdisciplinary adaptation and expansion of existing REC by integrating computer and data science into existing REC bodies, including experts from additional disciplines and training (Stahl et al., 2025; Brenneis et al., 2024) and introducing new principles and guidelines (Bouhouita-Guermech et al., 2023; Hagendorff, 2020). Apart from notions of such 'Super-REC', the additional creation of specialized sub-committees is advocated (Esmaili et al., 2025). Deviating from this idea of `traditional´ REC assuming additional tasks there is also a strand of literature that puts forth the idea additional separate and specialized REC, dubbed, e.g., as Algorithmic Research Ethics Committee/Board (ARB) or AI Research Committee (AIRC) (Hadley et al., 2024; Jordan, 2019).and including flexible approaches such as ETHNA (González-Esteban and Calvo, 2022).

Our analysis further builds on scholarship that explores the co-constructivist nature of techlaw, i.e. the idea that both, hard and soft law norms evolve together with technological development in mutual dependency (Jones, 2018; Crootof and Ard, 2021). This approach emphasizes the formative role of norms within innovation environments and offers a theoretical grounding for designing regulatory sandboxes as sites of iterative governance rather than merely reactive or pre-emptive control. This research also allows for an inclusion of the EU AI Act´s risk-based approach into the AI research governance (Resseguier and Ufert, 2024; Wernicke and Meding, 2025).

Since REC are questioned as the optimal structure for AI research ethics oversight (Stahl et al., 2025), we consider the growing body of scholarship that examines the emergence and operationalization of regulatory sandboxes as instruments for normative experimentation and 'moral imagination' (Undheim et al., 2022): This includes research on their role in the EU's AI Act, where regulatory sandboxes are proposed as controlled

environments for innovation under public supervision, focusing on their potential to balance innovation with regulatory oversight (Plato-Shinar and Godwin, 2025; Undheim et al., 2022; Ranchordas, 2021). While regulatory sandboxes have been primarily understood as tools for innovation governance at the state or market level (Gumbo and Chude-Okonkwo, 2025), we build on ethical approaches to regulatory sandbox conceptualizations (Francis, 2025) and propose an integration of responsible AI frameworks (Göllner et al., 2024a). Our approach can therefore be categorized as 'intra method', as it primarily deals with the responsible design of technology (Reijers et al. 2018).

3 Methodology

This article adopts a conceptual and analytical methodology, combining theoretical inquiry with normative analysis. The study draws on interdisciplinary frameworks, primarily from STS, responsible AI research and legal theory, to examine the underlying logic and implications of an AI research ethics sandbox based on responsible AI considerations. Through this approach, the paper aims to clarify key concepts, identify underlying assumptions, and develop a structured argument based on existing literature and normative reasoning. This approach is underpinned by the notion of reflexive governance (Voss & Kemp, 2006; Feindt et al., 2018), which emphasizes iterative policy development, stakeholder deliberation, and the institutionalization of uncertainty. Such a perspective is especially pertinent to the AI research domain, where the consequences of methodological and technical decisions are often opaque, distributed between research disciplines and stakeholders, and temporally deferred. To this methodological end, the paper employs analogical reasoning and governance learning (Stone, 2012; Sabel and Zeitlin, 2012; Rangoni, 2022) to identify structural and functional parallels between regulatory sandboxes in innovation policy and the exigencies of ethical review mechanisms in AI research governance. The approach involves the mapping of problem similarities (e.g., uncertainty, novelty, rapid change), institutional roles (e.g., regulatory gatekeepers, facilitators, enablers), and process characteristics (e.g., iterative evaluation, stakeholder feedback loops). This 'cross-pollination' with sandbox logic opens a space for reflexive governance (Voss & Kemp, 2006), wherein ethical oversight becomes an iterative and participatory process rather than a fixed ex ante assessment. By integrating these methodologies, we purpose a conceptual basis for rethinking how academic research governance can responsibly adapt to emerging challenges such as AI or algorithmic experimentation as well as data-intensive applications, transferring regulatory sandbox logic to AI research ethics oversight.

4 Pitfalls in the Ethical Governance of AI Research

4.1 The Role of Research Ethics Committees in Research on and with AI

REC are entrusted with safeguarding ethical standards and balancing academic freedom with societal responsibility. Their role has become increasingly complex in the context of AI research, where both the pace and epistemic configuration of research challenge the assumptions underpinning conventional review processes of REC (Gonzáles-Esteban and Calvo, 2022; Esmaili et al., 2025). The normative principle of *rule-bound academic freedom* demands that universities, while enjoying institutional autonomy, demonstrate responsibility in overseeing potentially harmful research (Wernick and Meding, 2025). REC embody this responsibility, yet their practices remain rooted in paradigms often not doing justice to the technical and systemic features of AI. AI research often implicates diffuse, systemic, unpredictable harms, whether to privacy, fairness, or fundamental rights, issues that exceed the scope of traditional ethics assessment (Jobin et al., 2019; Mittelstadt, 2022).

The governance challenges differ markedly between *research on AI*, i.e. research which targets AI as its object and usually involves algorithmic development, and *research with AI*, where, often proprietary, AI algorithms serve as tools for inquiry in other domains and disciplines (Stahl et al., 2025). The former entails direct engagement with the design, testing, and evaluation of AI systems, while the latter embeds AI within disciplinary contexts where its limitations and embedded values may remain obscured. REC are increasingly tasked with reviewing both types, often without tailored methodologies or a shared vocabulary of risk.

4.2 Structural Critique: Why the Traditional Ethics Review Procedure Falls Short in AI Research

Structured around linear application and approval procedures, the traditional ethics review model often is unsuitable for AI research. REC were designed to assess clearly scoped studies with stable methods and foreseeable risks. AI research, by contrast, often unfolds within iterative, interdisciplinary, and exploratory projects whose normative and epistemic contours emerge during the research process itself (Stahl et al., 2025). In many AI and big data projects, key ethical dimensions, especially fairness, trustworthiness, human-centeredness, bias mitigation, and explainability, are not predefined checkboxes but outcomes of ongoing technical, conceptual, and empirical work (Gonzáles-Esteban and Calvo, 2022). The same applies to fundamental rights implications, which are frequently discovered or clarified only through experimentation (Bouhouita-Guermech et al., 2023; Díaz-Rodríguez et al., 2023). This `epistemic fluidity´ produces a structural tension: REC are expected to conduct anticipatory assessments of research that involves algorithmic and non-algorithmic rules yet to be fully articulated. As a result, ethics

oversight risks devolving into a formalistic exercise, imposing rigid scrutiny on processes that require ongoing adaptive reflection, resulting in review gaps (Reijers et al., 2018; Zimmer, 2018). This structural dilemma is amplified by the AI Act's explicit exemption of academic research from its binding regulatory scope, leaving REC as de facto ethical gatekeepers. However, they are poorly equipped for this role without appropriate institutional tools or temporal flexibility. The principle of precaution itself is put at risk: premature or shallow review may inadvertently sanction a normative vacuum, with ethical reflection sidelined until after deployment or publication.

These challenges are not merely theoretical and touch upon almost all disciplines and modes of AI use (Brenneis and Burden, 2025; Masso et al., 2025). Issues arise, e.g., from proprietary software, opaque model behaviors, and the re-identification potential of anonymized datasets (Esmaili et al., 2025). Across diverse cases one pattern recurs: the intertwining of data and algorithmics generates ethical risks that extend far beyond data protection. These include proprietary model opacity, normatively charged classification decisions, and epistemic displacement, where complex social judgments are outsourced to probabilistic systems. Such risks cannot be adequately anticipated in advance because they emerge from the interplay of technical architecture, data provenance, and research context.

Moreover, when AI functions as an epistemic infrastructure rather than as an object of inquiry, its normative implications risk becoming invisible. This is particularly problematic in applied fields, where AI tools are operationalized without critical reflection on their embedded assumptions. In such fast-moving research environments, requiring formal review for every AI application is impractical and potentially stifling. What is needed instead is an adaptive, recursive governance model, i.e. oversight that enables parallel ethical reflection.

4.3 Co-Constructing Responsibility

Debates about the governance of emerging technologies are often framed around a reactive normative paradigm: hard law and soft law perpetually scrambling to catch up with the speed of innovation (Calo, 2015). This 'normative lag' narrative, emblematic of what Jones (2018) refers to as 'technological exceptionalism', positions normative frameworks as outdated or inert in the face of rapid technological transformation. But this framing misrepresents the complex reciprocity between technical and normative systems. The governance of AI research cannot be meaningfully addressed through a simple linear model in which (ethical) norms react to technological change. While it is often said that rules 'lag behind' innovation, this narrative of reactive governance obscures a more fundamental dynamic: technology and regulation are mutually constitutive (Jones, 2018; Kaminski, 2023). Norms do not merely constrain or respond to technological development; they are co-produced with it, shaping what is built, how it is tested, and what is ultimately seen as acceptable or desirable. According to this

scholarship, technology can be characterized as a *socio-legal construction*, a view that waives the idea of technology as a neutral or autonomous force and instead emphasizes the normative architectures in which it is embedded from the start. Norms are to be seen as an infrastructural component of innovation, involved in everything from institutional design to the allocation of liability and legitimacy (Crootof and Ard, 2021). Seen through this lens, rules are not merely a set of exogenous constraints: One of the defining characteristics of AI research is its experimental, iterative, and exploratory nature. Many projects do not begin with fixed hypotheses, stable methodologies, or clearly anticipated outcomes (Esmaili et al., 2025). Instead, they involve speculative inquiry into algorithmic behaviour, emergent model properties, or complex data interactions. As such, key normative categories such as trustworthiness, fairness, transparency, bias mitigation, human-centeredness, are not static benchmarks to be checked off but are formed and refined through the research process itself.

The co-constructive nature of AI research governance is relevant for both, *research on* and *research with* AI. In research *on* AI, normative concerns are often an explicit part of the development process, as researchers examine issues such as algorithmic bias, model robustness, or the implications of scale and generalization. In contrast, research *with* AI uses AI systems as tools or infrastructures that support inquiry in other disciplines. Whether used in medical diagnostics, nursing, or historical text analysis, AI becomes a background instrument. Yet precisely in these settings, the embedded assumptions, data dependencies, and potential harms of AI systems can become invisible (Masso et al., 2025). Normative issues, ranging from bias and opacity to re-identification risks, may go unexamined or at least not understood sufficiently because AI is seen not as a subject of scrutiny, but as a technical utility.

This ‘rules-in-the-making’ logic creates a structural challenge for traditional ethical governance mechanisms. REC, while essential for safeguarding rights and ensuring accountability, are often tasked with prospectively assessing projects against normative standards that are not yet clear. This creates a temporal and epistemic mismatch: ethics oversight mechanisms are expected to anticipate and evaluate risks that are still unfolding and often unknowable in advance. This is not a matter of regulatory failure but of structural incompatibility: AI research generates its own norms as it proceeds, particularly in domains such as fairness and explainability, where solutions are context-sensitive and often co-produced in dialogue with technical, disciplinary, and societal inputs. REC, designed around anticipatory governance, find themselves at the limits of their institutional design: asked to adjudicate the ethical soundness of research trajectories whose parameters are still under construction. A co-constructive understanding of AI research governance thus requires moving beyond linear or top-down models of oversight.

Responsibility is not a pre-defined standard to be enforced ex ante; it is an evolving, situated, and distributed practice. Researchers, technologists, legal experts, ethicists, and communities must, depending on the research, participate in shaping what responsibility means in context, i.e. across different phases of research, applications, and institutional settings (Stilgoe et al., 2020; König et al., 2021). This reframing also calls for rethinking the institutional role of ethics governance bodies. Rather than serving primarily as gatekeepers issuing once-off approvals, REC might serve their role better if they also facilitate ongoing normative reflection, and, in so doing, support researchers in articulating, contesting, and refining the values that guide their work (Masso et al., 2025). This requires institutional learning mechanisms, interdisciplinary dialogue, and openness to the provisional nature of ethical judgments in complex, high-uncertainty domains like AI. In short, norms do not merely follow technology, nor can ethics be 'applied' to research like a seal of approval. Both are part of the architecture of innovation itself. In responsible AI research, norms and systems must be developed together, in an iterative and participatory manner that acknowledges their mutual dependencies (Göllner et al., 2024b). Only by recognizing and institutionalizing this co-constructive dynamic governance frameworks can be developed that are genuinely capable of meeting the ethical and epistemic challenges of AI research.

4.4 Risk-based Approach to AI Research Ethics Assessment

With its risk-based framework, the EU AI Act offers a useful, albeit indirect, point of reference for soft law approaches (Gille et al., 2024). Though the AI Act exempts academic research covering AI systems and AI models, including their output, from its formal scope (AI Act, Art. 2(6), rec. 25), its emphasis on risk classification, harm mitigation, and fundamental rights protection provides REC with an external source of normative orientation (Resseguier and Ufert, 2024; Wernicke and Meding, 2025). Complementary frameworks, such as the European Commission's High-Level Expert Group (HLEG) on AI and its 'Ethics guidelines on trustworthy AI' (HLEG, 2019; Göllner et al, 2024a), offer further guidance for the formulation of internal ethics policies and procedural criteria.

AI research invariably operates under conditions of uncertainty, where risks, technical, ethical, social, and systemic, are often diffuse and emergent. In this landscape, legal and ethical standards cannot be cleanly codified in advance but must evolve with and through technological practice. At the same time, risk is not an anomaly to be eliminated but a constitutive feature of the policy choice underpinning risk-oriented regulation of digital innovation, a 'risk baggage' (Kaminski, 2023). This constitutive aspect demands a structural response: risk mitigation must be integrated into research processes, not appended after the fact, for the research is in many cases tested in real-world situations and is often likely to end up in applications in market environments. This aspect goes even further, namely in ethics-by-design methodologies ('EbD-AI'), i.e. the

comprehensive and systematic inclusion of normative-ethical considerations in the design and development of AI (d'Aquin et al., 2018; Brey and Dainow, 2023). Such ethics-informed design and development considerations would also bring the research output into line with notions of fundamental rights impact assessment (FRIA) brought in by the EU AI Act (Mantelero, 2024).

5 The Responsible Artificial Intelligence Sandbox

5.1 Responsible Artificial Intelligence in the Research Ethics Review Process

Not all AI research projects require formal approval by REC. A categorical distinction is necessary to preserve both the integrity of ethics governance and the operational feasibility of review processes. A differentiated approach, based on technical and epistemic considerations, enables a more targeted allocation of ethical oversight and is tentatively delineated as follows:

1. Research activities that involve human subjects, process sensitive personal data, are security-related, or have foreseeable implications for individual rights and social outcomes, must remain subject to the standard REC procedure. This includes studies deploying AI in projects, e.g., involving biometric recognition, behavioural prediction, or automated decision-making with high-stakes consequences. The same applies also to any research that evaluates or calibrates AI systems using personally identifiable or vulnerable data.
2. Purely technical or foundational research on AI models that does not involve human participants, personal data, or application scenarios with immediate ethical relevance can be excluded from mandatory REC review. This research includes algorithm development, benchmarking on synthetic or anonymized datasets, formal model analysis, or optimization/refinement studies where no direct or indirect harm is plausible. Requiring ethics review for such abstract work would prevent unnecessary procedural overhead.
3. Between these two categories lies a ‘grey zone’ of interdisciplinary and application-oriented AI research that operates under conditions of epistemic uncertainty and dynamic normative standards. Such research, e.g., using pretrained models in new domains, combining social datasets with machine/deep learning techniques, or deploying AI in exploratory decision support scenarios, often does not initially meet the criteria for formal REC review, yet may develop ethical risks over time.

Responsible AI sandboxes can address the issues of the third category and allow researchers to conduct reflexive assessments dynamically adjusting review thresholds.

This way a sandbox supports sensitivity analyses across different models and application domains, enabling the systematic ethics evaluation.

5.2 Responsible Artificial Intelligence Sandbox

5.2.1 A sandbox for Responsible Artificial Intelligence Research and Innovation

The concept of the sandbox, borrowed from software engineering and regulatory experimentation, offers a productive frame to resolve challenges outlined in the previous sections. Where technological sandboxes isolate code from production environments to allow for safe experimentation, regulatory sandboxes extend this notion to legal and ethical governance, providing time-bound, supervized environments for controlled testing under conditional flexibility (Ranchordas, 2021; Longo and Bagni, 2025). The concept of the regulatory sandbox, increasingly common in technology regulation, has gained traction as a model for iterative, adaptive governance under uncertainty. However, to address the distinctive complexities of AI research within academic settings, a further evolution is required: the development of a Research and Innovation Sandbox for responsible AI research.

Regulatory sandboxes provide a practical instantiation of the co-constructive logic outlined in section 4. Traditionally conceived as ‘safe spaces’ for testing new technologies under experimental regulatory supervision, sandboxes offer more than conditional regulatory relief. In the context of AI research, their real potential lies in enabling legal, ethical, policy and technological actors to engage in and pioneer real-time governance experiments. Far from being stopgaps for legal uncertainty, sandboxes institutionalize learning, both legal and ethical, by allowing situated, iterative norm formation together with technical development. Such regulatory AI sandbox reconfigures governance from a paradigm of procedural oversight to one of responsible epistemic co-production (Seferi, 2025). Rather than serving as a temporary regulatory exception, the sandbox becomes a continuous, institutionally embedded learning space for situated ethical inquiry and iterative norm development. It is conceived as a distributed and integrative research infrastructure, a ‘playground’ (Resnick, 2017) in the constructive sense, where researchers from different disciplines and domains, ethicists, legal scholars, and societal stakeholders can collaboratively engage with the uncertainties and contested values of AI systems (Undheim et al., 2022).

5.2.2 The Role of the Research Ethics Committee

The governance of the co-constructive research and innovation environment calls for a redefinition of the role of REC. Rather than bypassing REC, sandboxing invites a reconfiguration of the REC´s tasks: from reviewing individual AI experiments (risk category 3, section 5.1) to overseeing the sandbox as a meta-level governance framework. This model empowers REC to evaluate the quality, reflexivity, and

accountability mechanisms of the sandbox itself, without constraining research within premature or inadequate normative templates. An intra-university responsible AI sandbox could function like a specialized REC sub-committee. This shift supports a holistic, multidisciplinary perspective and concentrates AI expertise within a reflexive, iterative, and ethics-by-design research process environment. Serving as a gateway for external collaboration, the sandbox enables pre- and post-approval engagement while embedding discursive reflection within an established normative framework.

The internal university governance setup, comprising an REC and a responsible AI sandbox, can draw valuable insights from legal/regulatory sandbox models, despite the absence of standardization. Three distinct models emerge: a narrow model focused on product testing (e.g., automated driving); a broad model aimed at regulatory experimentation (e.g., frameworks for AI regulatory sandboxing, Art. 57 AI Act); and hybrid forms enabling context-specific testing of anticipated regulatory regimes, as seen in pre-AI Act initiatives in Spain (Bagni and Seferi, 2025). These configurations (Mobilio and Gianelli, 2025) offer design principles transferable to the internal university context: Most notably, the notion of an *explicit carve-out*, i.e. a defined and bounded space for experimental activity, can be mirrored in a sandbox environment sanctioned by the REC. Within this space, delegated authority enables ethical and regulatory experimentation through iterative and co-constructive processes. A sub-committee structure or designated oversight body could exercise *discretionary powers* akin to those of an oversight authority, facilitating context-sensitive governance of AI research. Establishing *risk thresholds* linked to ethical, technical, and societal dimensions permits differentiated levels of oversight, aligning with REC concerns while preserving room for innovation. *Evaluation and reporting* mechanisms institutionalize reflexivity and ensure accountability over time, supporting a shift from static approvals to dynamic governance.

Crucially, the university sandbox, like its regulatory counterparts, can function as a site of evidence-based regulatory learning. By creating structured input for internal governance and potentially informing external norms, the sandbox helps bridge experimental research and anticipatory regulation (Morgan, 2023). This allows the REC and responsible AI sandbox to evolve beyond compliance gatekeeping towards a model of anticipatory, participatory, and reflexive oversight. Drawing on the AI Act's guiding principles, as well as complementary frameworks such as the EU High-Level Expert Group's (HLEG) ethics guidelines for trustworthy AI (HLEG, 2019), we propose that sandbox environments operationalize 'responsibility' through embedded assessment criteria and metrics (Díaz-Rodríguez et al., 2023). We propose a conceptual structure for capturing these dimensions across the lifecycle of AI systems (section 5.3). To ensure the normative legitimacy of the RAIS, ethical governance must, of course, rest not only on open-ended deliberation but also on procedural clarity, inclusiveness, and transparent communication of expectations and assessment criteria. The AI Act with its risk- and

fundamental-rights-based approach and the HLEG´s guidelines can, at least in the EU, serve as a normative compass in this regard.

5.2.3 Operationalizing Responsible AI by Embedding Normative Reflexivity

The following operationalization of responsible AI for research ethics review purposes is combined with an analysis of regulatory sandboxes as tools for moral imagination (Undheim et al., 2023; Resseguier, 2024) and operates within a framework for responsible innovation, developed in the governance of emergent technologies (Stilgoe et al., 2020). By synthesizing these perspectives, we aim to devise a model for intra-university engagement with AI that is scientifically and pedagogically generative as well as ethically responsive and reflexively designed. Regulatory sandboxes serve a function beyond the minimization of (legal) risk or the facilitation of technological deployment but are valuable precisely because they enable moral imagination under conditions of 'true uncertainty', scenarios in which outcomes cannot be reliably predicted or calculated (Undheim et al., 2023). Within such spaces, the focus shifts from managing known risks to cultivating anticipatory and adaptive forms of governance. This orientation is particularly relevant in the context of AI, where systems increasingly interact with complex social environments and generate outcomes that escape straightforward evaluation. The collaborative, interdisciplinary, and iterative processes within sandboxes emphasize the role of co-learning among regulators, developers, and affected publics. While regulatory sandboxes are typically situated within governmental or market-facing institutions, the core insight that ethical AI development requires structured spaces for open-ended exploration (Francis, 2025) can be usefully transposed into the academic setting. A university-based responsible AI sandbox would not replicate the function of a regulatory sandbox in a narrow sense but would instead adapt its enabling logic to support transdisciplinary learning and responsible design practices. Such sandbox derives its legitimacy from its institutional role in enabling deliberation, capacity-building, and critical inquiry, making it well-suited to address the educational dimensions of responsible AI. By involving students and early-career researchers in real-world projects under conditions of structured uncertainty, the sandbox offers a hands-on context for cultivating ethical sensitivity, interdisciplinary literacy, and design reflexivity. Moreover, the sandbox can serve as a platform for developing soft-law instruments such as codes of conduct, evaluation protocols, or value-sensitive design guidelines that extend beyond individual projects and contribute to the institutional culture of responsible AI development.

The reflexive process can (and should) draw on approaches to algorithmic impact assessment (Selbst, 2021), such as the Fundamental Rights and Algorithms Impact Assessment (Gerards et al., 2022), which provide structured methodologies for identifying and mitigating human rights impacts throughout the algorithmic development process. Integrating such approaches into the framework and assessment practice

strengthens its ability to operationalize fundamental rights considerations, linking reflexive ethical deliberation with concrete, legally informed assessment practices. This alignment would also enhance the approach´s accountability dimension by situating ethical reflection within broader societal and legal frameworks of rights protection and participatory evaluation.

In addition to traditional ethical concerns, responsible AI governance must also address questions of sustainability, encompassing the environmental impact of computationally intensive methods, the infrastructural dependencies of AI research, and issues of digital sovereignty. Particularly in academic settings, where AI experimentation often relies on energy-intensive processes and third-party cloud infrastructures, ethical reflection should extend to the ecological footprint and long-term viability of research practices.

5.3 Approach to a Responsible AI Sandbox - a Focus on Metrics

The technical perspective of the operationalization of responsible AI within a sandbox environment requires a technically grounded and multidimensional algorithmic impact assessment framework. Its purpose is to enhance deliberation with technological means (Mauri et al., 2024). We propose a categorization of assessment dimensions that, at the same time, prepare the ground for ethics-by-design methodologies

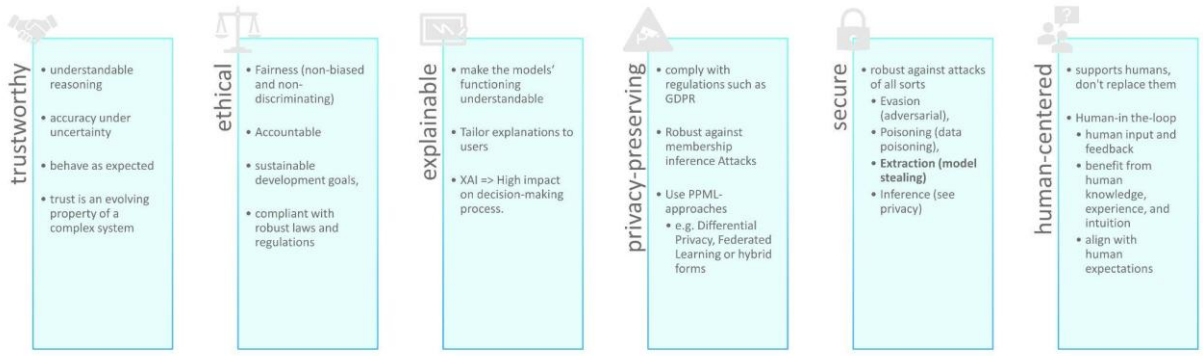


Figure 1: Responsible AI metrics categorization, based on Göllner et al., 2024b.

(Brey and Dainow, 2023) by enabling evaluation metrics in pillars as shown in Fig. 1 (based on Göllner et al., 2024b; High-Level Expert Group on AI, 2019). These pillars function as a guideline or normative-technical blueprint for design, evaluation, and governance of AI systems and embed responsible AI into research and innovation development:

Trustworthiness: AI systems should demonstrate understandable reasoning under uncertainty, maintain accuracy and functional reliability across varying input conditions, and behave consistently with expectations. Possible metrics under this category include calibration error, uncertainty quantification, and out-of-distribution detection rates. In sandbox settings, these metrics are evaluated continuously across model updates to

monitor trust degradation or improvement over time. Trust is not static but evolves as a system interacts with complex environments. It must be treated not as a binary property, but as a dynamic signal within the model lifecycle.

Ethical Alignment: The ethical (in the narrower sense) assessment focuses on fairness, non-discrimination, and accountability. Technical fairness metrics, such as demographic parity, equalized odds, and disparate impact, are computed over protected attributes. Bias mitigation strategies, such as reweighting, adversarial debiasing, or fair representation learning, can be integrated into the sandbox's evaluation logic. Accountability further demands traceability, supported, e.g., through lineage logging. This aligns with the Ethics-by-design approach, focusing on integration of ethical considerations into AI system development from the start (Brey and Dainow, 2024).

Explainability: The decision-making processes of AI systems should be interpretable. Explainable AI (XAI) methods can convey an idea about the decision-making process and support transparency of AI systems. For the proper interpretation of the explanations, domain/expert knowledge is required. Explainability is addressed through both post-hoc and model-intrinsic approaches. A sandbox should support post-hoc explainability via model-agnostic methods (e.g., LIME, SHAP) and model-specific techniques (e.g., Integrated Gradients). Quantitative evaluation dimensions include faithfulness (e.g., input perturbation tests), monotonicity (whether feature importances correlate with performance), sparsity, and explanation stability under perturbations. Explainability-by-design approaches, such as attention-based architectures or self-explaining models, should be benchmarked for interpretability scores. Domain knowledge is necessary for contextual evaluation of explanation plausibility.

Privacy Preservation: Robustness against privacy leakage is essential, especially when handling sensitive or personal data. Technical evaluation includes membership inference attacks, where the area under the ROC curve (Receiver Operating Characteristic) serves as an indicator for memorization risk. Privacy-preserving machine learning techniques, such as differential privacy, federated learning, or secure multiparty computation, must be incorporated where applicable. Privacy guarantees should be auditable, with the sandbox providing reproducible attack simulations and quantitative leakage indicators.

Safety and Security: AI systems should be resilient to a wide spectrum of attacks, including evasion (adversarial examples), data poisoning, model extraction, and inference. AI systems deployed in dynamic environments should be therefore robust to input perturbations. Security assessments include adversarial robustness, model extraction resilience, and poisoning attack tolerance. Robustness is measured under different threat models using standardized frameworks, employing metrics such as worst-case accuracy under bounded perturbations. Model extraction and inversion risk are assessed via black box querying strategies. Poisoning robustness involves training set sanitization efficacy and resilience of performance under perturbed training distributions.

Human-centeredness: Human-in/on/beyond-the-loop architectures ensure that human judgment, experience, and feedback remain integral, aligning AI behaviour with human expectations and societal values. This assessment pillar emphasizes the role of the human not only as an end-user but as an epistemic agent. Human-in-the-loop setups should be formally integrated into the model evaluation cycle, enabling structured user feedback loops, active learning setups, or override mechanisms. Metrics in this category include task performance with vs. without human correction, agreement scores between model and human judgment, and usability or cognitive load assessments. Models must align with human expectations and support decision augmentation, not replacement.

From the technical perspective, metrics-based evaluation is a necessary component for operationalizing responsible AI. Metrics are quantitative indicators which provide measurable criteria for the system behaviour. Yet quantitative metrics alone cannot fully capture the reliability and compliance across all dimensions of AI systems. Responsible AI involves context sensitivity, trade-offs (e.g., between transparency and security) and such aspects as trust or human-centeredness. Many decisions regarding AI depend on domain knowledge, interpretation or ethical judgment. Therefore, it is essential to engage experts and users not only to interpret metric results, but especially to identify limitations and make development decisions. The expert input should be based on interdisciplinary knowledge and integrated directly into the evaluation process. Metrics help structure the evaluation of AI output, behaviour and risks. They also support systematic review that can be used by ethics committees to assess AI systems (Jordan, 2019). The VERIFAI Framework (Göllner and Tropmann-Frick, 2023) implements a large part of the metrics for AI classification models and thereby provides a foundation for the further technical development of comprehensive responsibility assessments. Its modular structure and initial metric coverage enable systematic integration of fairness, explainability, robustness, and privacy evaluations into the model development lifecycle. Given the breadth and heterogeneity of responsibility dimensions, a single monolithic tool is insufficient. Instead, a framework suite approach is better suited, allowing the flexible combination of specialized modules to address domain-specific requirements and to adapt metric application across different AI system types and application contexts.

6 Limitations

The implementation of reflexive ethics governance models such as RAIS must be understood against the backdrop of institutional constraints, as many REC in academic settings remain poorly funded and often lack the interdisciplinary expertise required for AI-related review. To ensure the practical viability of such frameworks, sustainable resourcing will be essential, possibly including dedicated funding lines, specialized staff positions, and closer integration with existing research infrastructure and support units.

Without such structural reinforcement, there is a risk that RAIS-like mechanisms could inadvertently add procedural complexity rather than strengthening ethical reflexivity.

The effectiveness of a reflexive environment such as the RAIS depends not only on its design but also on its alignment with prevailing academic incentive structures, which often prioritize competition, productivity, and intellectual ownership over deliberation and collective reflection. These pressures, combined with hierarchical dynamics within research teams, can discourage open discussion of ethical challenges, particularly among junior researchers or those in precarious positions. To counteract such effects, RAIS-like initiatives should be embedded within institutional frameworks that promote open science, ensure protection for critical participation, and establish participatory governance mechanisms that empower all members of the research community to engage safely and meaningfully in ethical dialogue.

7 Conclusion and Outlook

This paper has identified key structural limitations and epistemic mismatches that constrain REC in their capacity to adequately oversee AI research, particularly where uncertainty, iteration, and socio-technical entanglement are central features. By analogical reasoning from regulatory sandbox models in EU and national jurisdictions and by drawing on a comprehensive responsible AI framework, we have conceptualized the *responsible AI sandbox* (RAIS) as an institutional innovation capable of embedding ethical and legal norms *within* the research process rather than applying them ex ante or ex post.

RAIS functions not merely as a procedural alternative but as a co-constructive governance environment, where responsibility, fairness, risk-mitigation, and trustworthiness are shaped through bounded experimentation, iterative feedback, and situated reflexivity. Unlike traditional front-loaded ethics review, the sandbox allows for dynamic norm development aligned with the unfolding nature of AI technologies. In this model, the REC shifts from a gatekeeping role to one of conditional delegation and continuous oversight, enabling ethics to travel with the research. Such a transformation reframes university governance as an anticipatory, learning-oriented infrastructure bridging innovation and accountability.

The responsible AI sandbox is a proposed institutional infrastructure that embeds ethical reflexivity into AI research and innovation processes within universities. It addresses a ‘grey zone’ of AI research that does not clearly fall under existing REC ethics approval procedures but nonetheless raises emergent normative concerns. A differentiated review approach distinguishes foundational AI research (which may be exempt from REC review) from applied or high-risk projects (which require standard oversight), with the

sandbox offering a governance solution for projects in between, as well as a place for continuous research project reflection after formal approval.

RAIS introduces a metrics-based framework for assessing responsible AI, structured around six pillars: trustworthiness, ethical alignment, explainability, privacy preservation, safety/security, and human-centeredness. These dimensions enable ongoing evaluation across the AI lifecycle using technical and normative indicators, such as calibration error, fairness metrics, privacy leakage risks, adversarial robustness, and interpretability scores. Further, the RAIS framework can serve as a site for integrating sustainability considerations into ethical deliberation, promoting awareness of resource use, infrastructural resilience, and institutional autonomy as essential dimensions of responsible research on and with AI.

Inspired by regulatory sandboxes in law and technology, RAIS operationalizes ‘ethical co-construction’ through iterative, multidisciplinary collaboration. Unlike traditional oversight focused on compliance, the sandbox supports reflexive, adaptive governance, where researchers, ethicists, legal scholars, and societal actors collaboratively shape norms in real-time. This positions the REC not as a gatekeeper but as an institutional steward monitoring the sandbox’s integrity, transparency, and learning capacity.

Rather than approving individual projects, an embedded REC sub-committee can grant environment-level approval, enabling supervised ethical experimentation within a bounded domain. RAIS thus functions as a site of anticipatory regulation and institutional learning, generating soft-law instruments (e.g., codes of conduct) and ethical infrastructures. It draws legitimacy from its capacity to enable and facilitate deliberation, build capacity, and cultivate interdisciplinary responsibility.

Acknowledgements

This research was made possible through funding by the German Federal Ministry of Research, Technology and Space (Bundesministerium für Forschung, Technologie und Raumfahrt) as part of the Responsible Advanced Intelligent Methodologies and Skills Lab (R-AIMS) project at Hamburg University of Applied Sciences. The authors gratefully acknowledge this support, which made the present contribution possible. The authors would also like to thank the reviewers for the valuable comments. The views expressed are solely those of the authors.

Conflicts of interest

The authors declare no conflicts of interest.

References

d'Aquin, M., Troullinou, P., O'Connor, N. E., Cullen, A., Faller, G., & Holden, L. (2018, December). Towards an' ethics by design' methodology for AI research projects. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (pp. 54-59).

Bagni F. and Seferi F. (eds.) (2025), Regulatory sandboxes for AI and Cybersecurity. Questions and answers for stakeholders. CINI's Cybersecurity National Lab. ISBN: 9788894137378.

Bouhouita-Guermech, S., Gogognon, P., & Bélisle-Pipon, J. C. (2023). Specific challenges posed by artificial intelligence in research ethics. Frontiers in artificial intelligence, 6, 1149082.

Brenneis, A., Gehring, P., & Lamadé, A. (2024). Zwischen fachlichen Standards und wilder Innovation: Zur Begutachtung von Big Data-und KI-Projekten in Forschungs-Ethikkommissionen. Ethik in der Medizin, 1-19.

Brenneis, A., & Burden, T. (2025). Meeting report: 'Challenges Posed by AI for the Work of Research Ethics Committees'. Conference, 2024, Hannover, DE. TATuP-Zeitschrift für Technikfolgenabschätzung in Theorie und Praxis, 34(1), 70-71.

Brey, P., & Dainow, B. (2024). Ethics by design for artificial intelligence. AI and Ethics, 4(4), 1265-1277.

Centre Responsible Digitality (ZEVEDI): Research Ethics for AI Research Projects. Guidelines to Support the Work of Ethics Committees at Universities, Darmstadt 2023, 19 pp.

Calo, R. (2015). Robotics and the Lessons of Cyberlaw. Calif. L. Rev., 103, 513.

Crootof, R., & Ard, B. J. (2020). Structuring techlaw. Harv. JL & Tech., 34, 347.

Feindt, P. H., & Weiland, S. (2018). Reflexive governance: exploring the concept and assessing its critical potential for sustainable development. Introduction to the special issue. Journal of Environmental Policy & Planning, 20(6), 661–674. https://doi.org/10.1080/1523908X.2018.1532562

Ferretti, A., Ienca, M., Sheehan, M., Blasimme, A., Dove, E. S., Farsides, B. & Vayena, E. (2021). Ethics review of big data research: What should stay and what should be reformed?. BMC medical ethics, 22(1), 51.

Francis, K. (2025). The need for an ethical approach to regulatory sandboxes. In: Bagni F. and Seferi F. (eds.) (2025), Regulatory sandboxes for AI and Cybersecurity. Questions and answers for stakeholders. CINI's Cybersecurity National Lab.

Gerards, J., Schäfer, M. T., Muis, I., & Vankan, A. (2022). Fundamental rights and algorithms impact assessment (fraia).

Gille, M. & Tropmann-Frick, M. & Schomacker, T. (2024). Balancing public interest, fundamental rights, and innovation: The EU's governance model for non-high-risk AI systems. Internet Policy Review, 13(3).

Göllner, S., Tropmann-Frick, M., & Brumen, B. (2024a). Towards a Definition of a Responsible Artificial Intelligence. In Information Modelling and Knowledge Bases XXXV (pp. 40-56). IOS Press.

Goellner, S., Tropmann-Frick, M., & Brumen, B. (2024b). Responsible Artificial Intelligence: A Structured Literature Review. arXiv preprint arXiv:2403.06910.

Göllner, S., & Tropmann-Frick, M. (2023). VERIFAI-A Step Towards Evaluating the Responsibility of AI-Systems. In BTW 2023 (pp. 933-941).

González-Esteban, E. & Patrici, C. (2022). Ethically governing artificial intelligence in the field of scien.fic research and innovation. Heliyon, 8.

Hadley, E., Blatecky, A. & Comfort, M. Investigating algorithm review boards for organizational responsible artificial intelligence governance. AI Ethics 5, 2485–2495 (2025). https://doi.org/10.1007/s43681-024-00574-8

Hagendorff T (2020) The ethics of AI ethics. An evaluation of guidelines. Minds and Machines 30: 99–120. Crossref.

High-Level Expert Group on AI (HLEG) (2019) Ethics guidelines for trustworthy AI. Available at: https://ec.europa.eu/futurium/en/ai-alliance-consultation.1.html.

Jordan, S.R. (2019). Designing Artificial Intelligence Review Boards: Creating Risk Metrics for Review of AI, 2019 IEEE International Symposium on Technology and Society (ISTAS), 2019, pp. 1-7, doi: 10.1109/ISTAS48451.2019.8937942.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature machine intelligence, 1(9), 389-399.

Jones, M. L. (2018). Does technology drive law? The dilemma of technological exceptionalism in cyberlaw. U. Ill. JL Tech. & Pol'y, 249.

König, H., Baumann, M. F., & Coenen, C. (2021). Emerging technologies and innovation—hopes for and obstacles to inclusive societal co-construction. Sustainability, 13(23), 13197.

Lenzini, G. (2025). Artificial Intelligence Ethics: Challenges for a Computer Science Ethics Board with a Focus on Autonomy. In The Routledge Handbook of Artificial Intelligence and International Relations (pp. 382-391). Routledge.

Mantelero, A. (2024). The Fundamental Rights Impact Assessment (FRIA) in the AI Act: Roots, legal obligations and key elements for a model template. Computer Law & Security Review, 54, 106020.

Masso, A., Gerassimenko, J., Kasapoglu, T., & Beilmann, M. (2025). Research Ethics Committees as Knowledge Gatekeepers: The Impact of Emerging Technologies on Social Science Research. Journal of Responsible Technology, 100112.

Mauri, A., Hsu, Y. C., Verma, H., Tocchetti, A., Brambilla, M., & Bozzon, A. (2024). Policy Sandboxing: Empathy As An Enabler Towards Inclusive Policy-Making. Proceedings of the ACM on Human-Computer Interaction, 8(CSCW2), 1-42.

Mittelstadt, B. D. (2022). Principles alone cannot guarantee ethical AI. Nature Machine Intelligence, 4(1), 5–7.

Mobilio, G. and Gianelli, M. (2025). In: Bagni F. and Seferi F. (eds.) (2025), Regulatory sandboxes for AI and Cybersecurity. Questions and answers for stakeholders. CINI's Cybersecurity National Lab.

Morgan, D. (2023, August). Anticipatory regulatory instruments for ai systems: A comparative study of regulatory sandbox schemes. In Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society (pp. 980-981).

Petermann, M., Tempini, N., Kherroubi Garcia, I., Whitaker, K., & Strait, A. (2022). Looking before we leap: Expanding ethical review processes for AI and data science research.

Plato-Shinar, R., & Godwin, A. (2025). Regulatory Cooperation in AI Sandboxes: Insights from Fintech.

Ranchordás, S. (2021). Experimental Regulations for AI: Sandboxes for Morals and Mores. Morals & Machines, 1(1), 86-100.

Rangoni, B. (2022). Experimentalist governance. In Handbook on Theories of Governance (pp. 592-603). Edward Elgar Publishing.

Resseguier, A. and Ufert, F. (2023). AI research ethics is in its infancy: the EU's AI Act can make it a grown-up. Research Ethics, 20(2), 143-155. https://doi.org/10.1177/17470161231220946 (Original work published 2024)

Resseguier, A. (2024). Research ethics frameworks for artificial intelligence: The twofold need for compliance requirements and for an open process of reflection and attention. In Smart Ethics in the Digital World: Proceedings of the ETHICOMP 2024. 21th International Conference on the Ethical and Social Impacts of ICT (pp. 122-124). Universidad de La Rioja.

Reijers, W., Wright, D., Brey, P., Weber, K., Rodrigues, R., O'Sullivan, D., & Gordijn, B. (2018). Methods for practising ethics in research and innovation: A literature review, critical analysis and recommendations. Science and engineering ethics, 24, 1437-1481.

Resnick, M. (2017). Lifelong kindergarten: Cultivating creativity through projects, passion, peers, and play. MIT press.

Sabel, C. F., & Zeitlin, J. (2012). Experimentalist governance. In Levi-Faur, D. (Ed.), Oxford Handbook of Governance. Oxford University Press.

Seferi, F. (2025). A comparative analysis of regulatory sandboxes from selected use cases: Insights from recurring operational practices. In Regulatory sandboxes for AI and Cybersecurity. Questions and answers for stakeholders (pp. 145-176). CINI's Cybersecurity National Lab.

Selbst, A. D. (2021). An institutional view of algorithmic impact assessments. Harv. JL & Tech., 35, 117.

Stone, D. (2012). Transfer and translation of policy. Policy studies, 33(6), 483-499.

Voß, J. P., & Kemp, R. (2006). Sustainability and reflexive government: introduction. In Reflexive governance for sustainable development. Edward Elgar Publishing.

Zimmer, M. (2018). Addressing conceptual gaps in big data research ethics: An application of contextual integrity. Social Media+ Society, 4(2), 2056305118768300.

A new approach to sustainable development and decarbonisation of airport and seaport territories trough citizen science – HubCities

Sanela Pansinger, Tomaž Berčič

University of Ljubljana, Slovenia

DOI 10.3217/978-3-99161-062-5-011, CC BY 4.0

Abstract. The HubCities initiative addresses, in the context of other objectives, is the decarbonization of high-carbon airport and port areas as key hubs of European infrastructure. Ecological, social, and spatial factors are taken into consideration in the project, along with the involvement of the local population. Participatory planning and citizen science play a central role in this process. Therefore it is essential to underscore that citizens take an active role in identifying, planning, and implementing new measures. In this manner the project works towards local transformation via the combination of CO_2 reduction and integration with social justice with aim to reach more sustainable spatial organizations. HubCities as a result acquires valuable experience in the design of climate-neutral, human-scale future cities.

1. Introduction

Decarbonization of energy-demanding infrastructure and industry is one of Europe's greatest challenges in reaching its climate goals (European Commission, 2019; IPCC, 2021). Seaports and airports are especially significant in this context. As gateways to global trade, freight, and transport, they are of high economic importance. At the same time, they also emit massive levels of greenhouse gas emissions, consume numerous resources and they represent areas with a high ecological footprint (Graham and Marvin, 2001). However these places are not autonomous. They are situated in an extensive network that flows together with global supply chains, labor markets, settlement patterns, and mobility systems. Accordingly, in in the process of imagining new climate-neutral futures for airports and seaports, their connection to the surrounding environment must be considered just as much as their embedding within

regional and global contexts. Therefore, the approach requires an integrated framework that goes beyond purely technological reconfiguration, since technical measures alone have proven insufficient to achieve sustainable transformation. For that purpose, this strategy needs an integrated model that is not technologically reconfigured simply, as these measures alone have proved to be inadequate. Therefore, the European research

project ‘HubCities’ intends to establish new visions for the revitalization of these quarters as part of more sustainable urban and regional development. A ‘HubCities’ seeks to be recognised as integral component of a local fabric, shaped by needs, demands and habits by people who work, live and spend their time there (e.g. **Fig. 1**).

Fig 1: From left to right: Spatial organisation of the HubCities (yellow rectangle) around Graz Airport (AT), around the seaport of Koper (SLO), around Trieste Airport and around the seaport of Trieste (IT). Source: Google Maps, Graphic: S.Pansinger

The article introduces the HubCities project strategy. It describes strategic concepts for climate-neutral transformation developed across three European locations: Graz Airport (Graz, Austria), Koper Port (Koper, Slovenia), Trieste Port and Monfalcone Airport (Trieste, Italy). Special attention in the project is given to the local community's participation as active co-creators of transformation. Therefore HubCities places citizen science at its core. Citizen science in HubCities simultaneously offers valuable strategies for a ‘bottom-up’ transformation by integrating every-day knowledge as recognised equally relevant to planning and innovation as technical or scientific expert knowledge. This approach has been granted by the European Commission under the Seal of Excellence 2023 for the HubCities project. This initiative opens a broad spectrum of new possibilities for new infrastructure planning. Instead of further utilizing closed ‘transport machines’ as mono-functional devices, airports and seaports become living spaces that integrate work, recreation, production, mobility, and innovation. By this, previously closed-off areas become open and resilient climate-neutral transformation hubs. Citizen science is central to this transition by providing local communities with room for action, co-decision, and co-design. This helps to elaborate guiding visions for the transformation of these places that are rooted in local experience and need rather than being imposed as external ‘top-down’ solutions.

2. Theoretical framework

2.1 Transformation as a multi-scaler phenomenon

The transformation of seaport and airports into climate-neutral, resilient, and liveable spaces must be understood as a multi-scalar phenomenon. This phenomenon includes technological innovation, spatial and urban transformation, organizational innovation, and social change (Graham & Marvin, 2001). Therefore, this article is based on intellectual logic that combines numerous perspectives: the theory of citizen science, the theory of participatory transformation in planning and the vision for leading the HubCity as a multi-scalar, relational spatial configuration.

2.2 Citizen Science as a participation tool

In contemporary context, citizen science has become a central concept in academic research, spatial planning and governance (Hecker et al., 2018; Bonney et al., 2014). This methodology debates about the involvement of citizens in scientific initiatives – starting from the gathering of data and building research questions and hypotheses through to analysis and application of new concepts. In spatial, environmental, and climate science, this approach provides valuable insights into transdisciplinary research and practice. Citizen science examines how local experiences and knowledge have to be considered equally in order to develop more intelligent, practical, and sustainable have potential to answer to questions of complex spatial challenges . This interactive methodology has a number of advantages for planning: it incorporates realistic everyday knowledge regarding available resources, conflicts, needs, and direction setting on guiding visions. The most crucial is that supports trust building and acceptance of new ideas among the citizens. Despite the fact that supports ownership of one's own world, it enables forms of participation that go beyond information or consultation.

Despite many advantages, the inclusiveness of citizen science faces several limitations. The application of citizen science differs by region, citizens knowledge and socioeconomic background. Citizens with limited access to technology, language barriers or less participation awareness can be unintentionally excluded. The adoption of citizen science, as before mentioned, is also strongly dependent on social and spatial context. Participation opportunities are more likely to be available in urban areas, while rural or marginalized groups may face infrastructural or logistical barriers. There is also a real risk that citizen science would be used as an instrument to legitimize already made decisions, instead of truly empowering people to shape outcomes. When participatory processes are introduced at the late stage of planning or without open backchannel feedback loops, they are likely to be deployed as tools for persuasion, not co-creation. Behind this, the theoretical framing of citizen science often assumes that participation automatically leads to empowerment and equality. The assumption can reproduce

existing hierarchies and exclusions without active attention to power relations, however digital platforms can both enable and restrict participation. They enhance communication and transparency but can simultaneously reinforce inequalities based on digital literacy, access to devices, or language barriers.

2.3 Participatory planning as a transformational guiding principle

Citizen science must be grounded in a participatory planning approach in order that 'everyday knowledge' can be actualized and make a useful contribution toward the strategic action involved in the transformation of port and airport sites (Forester, 1999; Healey, 2006). Participation accordingly goes beyond the dissemination of information, this implies that has to be addressed as an equal component of a discourse where diverse perspectives are negotiated. Under these circumstances, the negotiating space, the 'agora' of public argument, offers a stage for expressing technical requirements, economic interests, and social visions (Habermas, 1996).

As a result, citizen participation in planning in the HubCities initiative required the incorporation of the following key aspects:

1. Early Involvement. Citizens, firms, public authorities, and civil society are engaged on board into planning processes from the beginning.
2. Open Communication. Framework conditions, guiding objectives, limits, and scenarios are readily provided.
3. Co-Creation. Guiding visions, scenarios, and concepts are co-created rather than imposed 'top-down.'
4. Negotiation. Contrasting ideas are negotiated, compromise solutions are designed, and strategic decisions are made by discussion.

2.4 HubCites as a theory of multi-scalar structure

This theoretical approach is grounded on the conceptual structure of the HubCities, which identifies the port and airport space as a multi-scalar, relational, and functional system. The hubs are usually presented as mono-functional 'transport node' or a closed spaces that are disconnected from its context. Rather than presenting in this way, this hubs need to be understood as local, regional, European, and occasionally global network of spaces, that shape or are shaped by the spatial systems to which they belong. With aim to develop deeper understanding these multi-scalar relations, this article conceptualizes hubs as relational nodes.

For instance:

1. The Port of Trieste is a local labor market, a European supply chain, and an international logistics network.

2. The Airport of Monfalcone plays a key role in local tourism, international connectivity, and the European strategy for transport decarbonization.
3. The Port of Koper acts as a gateway to Central Europe.
4. The Graz Airport is equally important for its surrounding region and for international connections.

While the multi-scalar framework demonstrates how hubs operate ad different spatial and functional scales, an STS perspective extends this view by questioning the technical processes behind these relations. It invites everyone to look beyond spatial scales of interactions between technology, society and power that align with operation these infrastructures. In this context, the multi-scalar shape of HubCities is not only spatial phenomenon but also technological construct, expression of broader institutional decisions, economic agendas and scientific compromises. HubCities can be also understood as an attempt to open ‘black box’ of technological neutrality while threating airports and seaport as socio-technical arenas where decisions about design, energy use and spatial development are shaped by citizens who live or work in this environment. With this perspective, HubCities project situates decarbonization not as optimisation process but as negotiation between different actors, values and scales that define what ‘sustainability’ actually is in practice.

2.5 Summary

The theoretical framework outlined before provides the foundation for developing transformational strategies at airport and port locations. Participation, citizen science, and the conceptual model of the HubCities collectively constitute the theoretical foundation for:

- understanding and coordinating multi-scalar interrelation;
- connection of experiential, daily and expert knowledge into planning process;
- develop climate-neutral transformational visions out of local participation;
- fostering innovations and test new ideas under real conditions;
- and reimagining current places into ‘living nodes’ for sustainable and resilient transformation.

This long term holistic strategy supports the process of decarbonisation (e.g. **Fig. 2**).

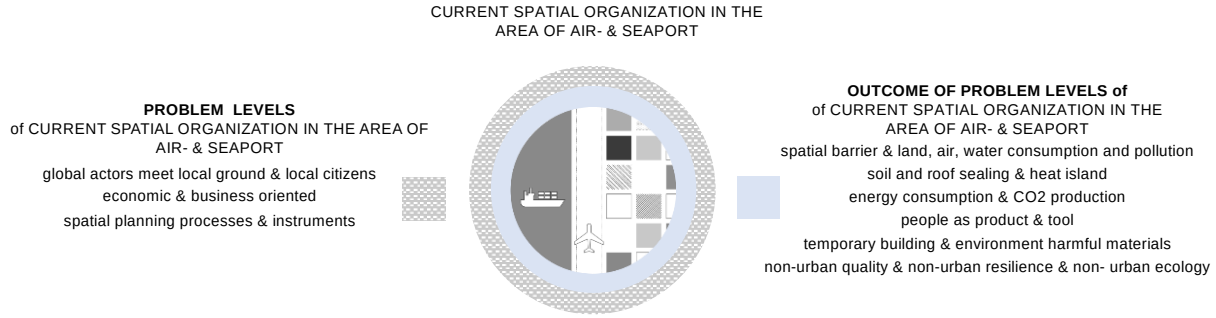


Fig 2: Current problem level in the field of airport and seaport areas. Graphic: S.Pansinger

3. Methodological Approach

The methodological approach adopted in this research project is integrative, transdisciplinary in nature that employs both qualitative and quantitative research methods with aim to determine the complex spatial, social, and cultural processes of port and airport spaces. The biggest aim is that transformation of such spaces is interpreted by considering the infrastructural, societal, and cultural drivers comprehensively.

The methodological framework is grounded of three main pillars:

1. Field survey and case studies in Graz, Koper, and Trieste cities
2. Participatory approaches and citizen science
3. Comparative analysis and transfer potential identification

3.1 Case studies based on field research

The empirical ground is based on qualitative case studies performed in the three selected cities: Graz (Austria), Koper (Slovenia), and Trieste (Italy). The cities are exemplary representations of varied urban and historic environments and thus are worthy of a comparative analysis of airport and port territories. The case studies are theorized as paradigmatic cases intended to identify commonalities and local differences.

The airport in the south of Graz, approximately 9 km from the city center, largely in the town of Abtissendorf in the town of Feldkirchen and partly in the cadastral town of Thalerhof in the town of Kalsdorf. Spatial embedding of the Graz airport in the wider

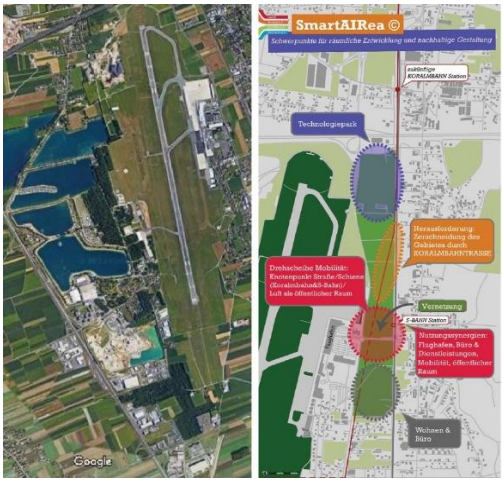


Fig 3: Current spatial organisation around Graz and smartAirea - www.smartairea.eu. One of the results shows that a polycentric development or the activation of the individual spatial areas of the airport environment offer the possibility of securing the spatial quality of the airport environment and thus at the same time the airport location in the urban-rural dimension. The challenges lie in a responsible process design that takes into account not only ecological, economic and social aspects, but also design and spatial aspects. Source: Sanela Pansinger, SmartAirea, GoogleMap

spatial context of housing, quarters, towns, region, and country is determinant for its overall beneficial contribution to spatial and urban development. Previously isolated regions around the airport are proposed to be integrated through manifold activities (life and work, transport and communication, food and entertainment). Graz Airport is designed to be a central node for international connectivity and an element of the city and region's mobility network. The link from Graz main station to the airport is especially highlighted as an axis of innovation, economy, and mobility. Alongside former railway lines, greenways, cycling routes, and footpaths have been developed to link the new quarter with the city center and surrounding residential areas. Industrial heritage is preserved and reinterpreted as part of new urban typologies. An area once exclusively logistical has become revitalized as a dynamic, multifunctional community with strong local identity potential (e.g. **Fig. 3**). Qualitative research methods were applied to the case study: in-depth interviews with local stakeholders from municipal administration, business community, and civil society, complemented by participatory on-site observations. Interviews focused on expectations, usage patterns, and perceived potential of the airport surroundings.

The Port of Koper, as the country's primary seaport and an interior inland transshipment center, is complemented strategically by the nearby Portorož Airport, enhancing regional connectivity and tourist accessibility. The use of the ‘axis-node model’ encourages more integration of the port and historic city center with enhanced passenger as well as freight transport through a newly formed corridor. The airport is a complementary transportation node that complements the international connectivity and maritime economy of Koper. The functional and physical proximity of the airport and port generates synergies in logistics, tourism, and innovation activities. As part of redevelopment city’s port efforts, numerous locations have been opened up to the public, such as promenades, parks, and restaurants, turning parts of the industrial landscape into a lively and accessible city zone. Meanwhile, formerly abandoned port buildings have been reconfigured to accommodate new and innovative uses, such as co-work, innovation hubs, and cultural facilities, thereby rendering the port itself a pulsating center of economic activity as much as urban

renewal. During the time the Port of Koper has evolved from a mono-functional economic space to a multifunctional space that brings together tradition, culture, and innovation. The positioning of Portorož Airport also helps this move forward by enhancing regional and global connectivity and bringing new economic dynamism into the master plan of the region (e.g. **Fig. 4**).

Fig 4: Professional guidelines for the Master Plan for the port of Koper, project team and source: Ažman, Venturi, Bercic et. al.

Being an economically significant port city, Koper provided a given environment under which the methodological focus was on the integration of participatory data collection and spatial analysis. Spatial trends of the port and its connection to adjacent neighborhoods were systematically mapped, with particular focus on transport corridors, green infrastructure, and accessibility. At the same time, focus groups were conducted with representatives of urban planning, ports authorities, and environmental NGOs to identify key planning questions, specifically ecological sustainability concerns and social inclusion concerns. The synthesis of spatial analysis findings and stakeholder deliberations enabled a better grasp of spatial interconnection and identified areas where optimization is possible within the port-city context.

As a former Habsburg port on the Adriatic, Trieste is noted for its high-density and multi-strand architectural and cultural tradition. The location of port, railroad, and airport facilities at their union produces a twin transport and economic node that is still framing urban and regional development. Trieste was a major city as a leading and prosperous port city in the 19th century, but its importance waned in the 20th century to make it one of the remaining port cities of the northern Adriatic (Ažman Momirski, 2021). As a result of increasing global competition within port logistics, Trieste has been continuously making efforts towards modernization and restyling. The city's Port Master Plan has been revised 24 times since 1957, the last time in 2010. The Free Port of Trieste is now divided into five diversified areas: three for commercial purposes (old free zone, new free zone, and timber terminal) and two for industrial purposes (mineral oils free zone and zaule channel free zone). The port follows the coastline, sea to city, but is spatially starved

through a shortage of storage and cargo space. Historically, large areas of valuable urban land have been used for port purposes (see Figure 10), excluding other possibilities for the city. In order to overcome these obstacles and unleash new urban opportunities, numerous strategic interventions have occurred. The central railway station has been reconverted into a multifunctional transport hub, fulfilling local, regional, and international mobility needs as well as business, educational, and gastronomic needs. Simultaneously, the newly built ‘activity corridor’ functionally and physically connects the port with the former old city center, enhancing the permeability of the city and enabling new uses that combine industry, heritage, and public life. Trieste Airport is central to this infrastructure chain by supporting maritime transport and reinforcing the city's status as a regionally significant and internationally connected node (e.g. **Fig. 5**). The city has further pursued a policy of cultural conservation and adaptive reuse. Previous port structures i.e., customs houses and warehouses have been reutilized as galleries, studios, and co-workspaces, thoughtfully conserving the cultural nature of the city while fulfilling modern urban functions. Cumulatively, Trieste's port, railroad, and airport complex is an officially designated urban and regional node. The planned convergence of old and new functions in old structures constitutes a hybrid spatial texture, one that intertwines through cultural heritage, economic innovation, and urban quality of life. Thus, Trieste is a model instance of sustainable and adaptive urban rehabilitation in the context of global change. The methodological approach employed in Trieste reflects the city's unique cultural and infrastructural legacy. Walk-throughs and systematic surveys were conducted in order to intercept both the continuity and the discontinuity of port–city development. In addition, ethnographic method, as participant observation at cultural events and interviews with residents and tourism industry and cultural arena actors were employed. This two-pronged approach allowed for a penetrating view of how the reshaping of infrastructure shapes local narratives, identities, and urban routine.

Fig 5: Spatial Organisation of the seaport of Trieste, and spatial organisation around the airport Ronchi in Trieste. Source: GoogleMaps

3.2 Participatory methods and citizen science

One of the central components of the strategy is the involvement of the local public and stakeholders interested in participatory formats. This is achieved through workshops, co-design events, and web-based platforms where the users are able to contribute their knowledge, needs, and visions. The purpose is to make planning processes more democratic, transparent, and user-oriented. One of the most important pieces of digital infrastructure supporting this strategy is the multilingual online platform HubCities (https://www.hubcities.net/), which has four language versions (e.g. **Fig. 6**). HubCities is an open-source platform that combines citizen participation, data visualization, and collaborative planning into a single digital platform. On the platform, it is possible to upload one's own observations or propose the development of urban and infrastructural spaces. The platform is already operational in all three case studies and facilitates the integration of virtual and analogue modes of participation (e.g. **Fig. 7**). The workshops face-to-face were each organized in tight cooperation with local partners in order to be able to respond to cultural and social specificities appropriately.

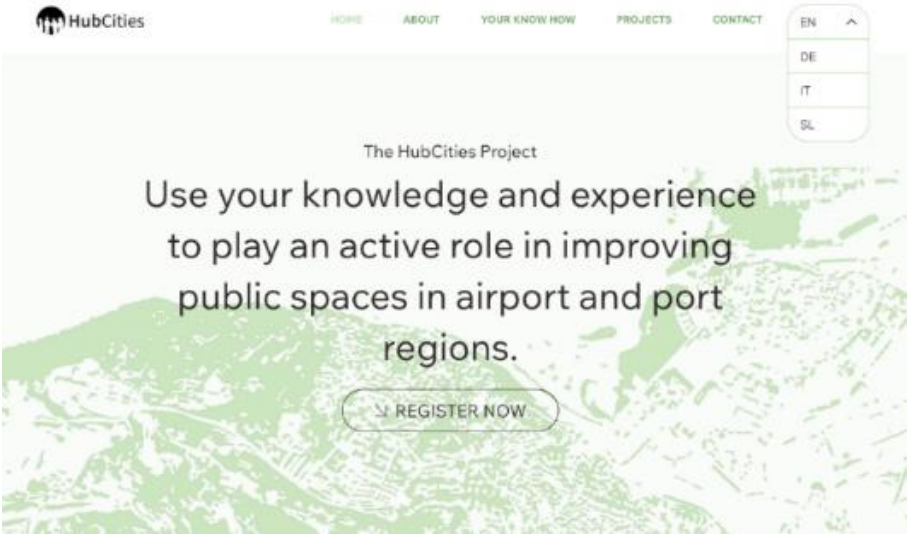


Fig 6. Online-platform HubCities, Graphics: www.hubcities.net

3.3 Comparative analysis and transfer potential

Fig 7. Announcement of workshop in Feldkirchen. Source: https://www.feldkirchen-graz.at/index.php/9-news/608-auftakt-workshop-des-projekts-hubcities,visited 15.06.2025.

Comparative analysis is the third methodology step, where results for Graz, Koper, and Trieste are compared to one another. The comparison shows both general tendencies and local-specific characteristics. The comparative analysis allows one to make general conclusions about sustainable development in port and airport regions outside the individual case studies. A particular focus is put on cultural identity as a basis for social cohesion and as an urban development resource in innovation. The findings demonstrate how infrastructure has not only a functional role but also symbolic significance for the residents, something which can be strategically utilized in planning processes. The lessons learned act as the foundation for guidance that can be applied to other comparable regions with consideration of the specific local context. The HubCities platform makes this possible through enabling the gathering and sharing of local experience and data in an international network thus ensuring effective transfer of knowledge. The combining of: qualitative field studies, participatory public engagement using digital and direct channels, and comparative analysis makes for a full comprehension of the complex dynamics in port and airport area development. By employing modern tools like the HubCities platform, the citizen science dimension is significantly improved, making planning processes more transparent, inclusive, and sustainable.

4. Results

Evaluation of questionnaires and participatory workshops within the three case study regions : Graz, Koper, and Trieste provides differentiated analyses of the potentials and challenges of sustainable development within airport and seaport areas on a local scale. The findings demonstrate the close relationship between ecological, social, and spatial aspects, and highlight the key function active citizen participation (Citizen Science) can assume in such processes. This case studies also identify the capability of online platforms like www.hubcities.net to enable experience sharing, knowledge sharing, and collaborative planning.

Graz-Thalerhof Airport in Graz was chosen for examination with a view to its integration within the urban setting. High environmental awareness of the populace found expression in noise and air pollution issues and general calls for decreases in emissions (e.g. **Fig. 8**).

Key findings of the workshops and surveys in Graz are:

1. Residents highlighted the importance of green and park spaces within close proximity to the airport. The spaces are not only considered as biodiversity providers, but also social places that enhance quality of living.
2. The improvement of living conditions in the immediate vicinity of the airport was perceived as a compensatory need for the infrastructure burden.
3. Transparency in dealing with environmental and climate information was repeatedly requested. Immediate emission data was requested to be shared through technological mediums by respondents, who viewed this as essential to the development of public trust among the public, authorities, and airport operators.
4. Involvement of citizens in collecting and interpreting data (citizen science) was identified to increase awareness of technical and planning procedures, and provide greater scope for acceptance of expected changes.

Fig 8. Comparing the current landscape with future scenarios without protection reveals risks of environmental degradation and loss of cultural identity. Graphic: S.Pansinger

Daily break and workplace interaction participants respond better to working conditions within the HubCities region. For lunch break, 47% of them had less than an hour, 18% had 1 to 1.5 hours, 12% had less than 15 minutes, and 23% indicated that they never have a lunch break. Lunch break variations time and workplace interaction demonstrate the degree to which time regimes and conditions at work shape individuals' capacity to participate in citizen science activity. Citizens employed in manufacturing sectors have less opportunity to participate in different activities, demonstrating clearly that participation is mirror of pattern and economic circumstances.

Inquiring about contact with employees of other organizations or residents, 44% reported no contact at all, 22% for less than 15 minutes, 28% for less than one hour, and only 6% for as long as 1.5 hours. This minimal organizational and community contact can be explained as both a consequence of spatial separation and the lack of shared social or collaborative space.

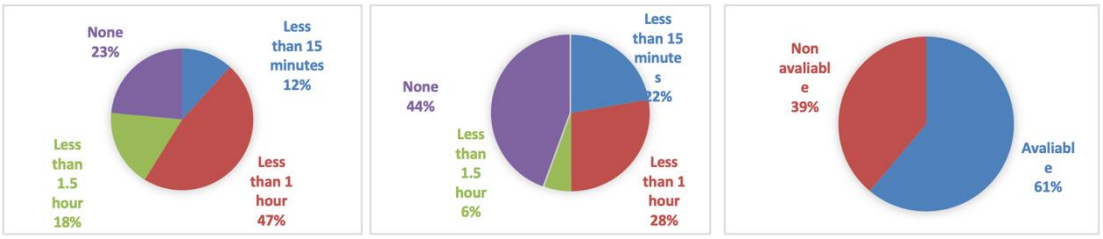


Fig 9. Results of survey analysis - Graz. Graphic: S.Pansinger

To a more positive aspect, 61% of the respondents have confirmed that recreational facilities such as gyms, swimming pools, or sporting fields are available within the HubCities area, while 39% have responded that there are no such facilities. It implies that, even though recreational facilities exist, they may be unevenly located or not fully utilized (e.g. **Fig. 9**).

Similarly, public and green space attitudes reflect socio-economic disparities. Higher availability of free time or cultural capital leads to putting environmental benefits in a different order than for groups whose everyday environment is industrial or logistical in character. It becomes evident from these findings that spatial preference and participation levels are socially organized and reliant on specific urban experiences.

The port city of Koper is defined by a tight economic interdependence with its port facility, a major employer and economic generator for the region. Surveying revealed that while the population is aware of the environmental impacts brought about by the port, its economic importance is strongly appreciated and recognized. The public at large endorsed better spatial integration of the port into the urban tissue. Stakeholders called for multifunctional land use concepts connecting ecological enhancement with social needs such as the creation of public green and leisure spaces that can also function as buffer zones between industry and housing. Participatory processes were viewed as necessary tools to balance complex and perhaps conflicting demands and translate them

into feasible and acceptable planning concepts. The use of the HubCities platform was deemed highly useful by the participants of the workshop in facilitating information flows and enabling early and ongoing citizen participation. The utilization of digital tools was not only applauded for improved communication but also for enabling the easy and clear visualization of emissions, traffic, and environmental impacts (e.g. **Fig.11**).

The findings indicate that 77% of respondents never go to public spaces such as parks or coffee houses during their breaks, indicating limited access to these spaces in daily activities most likely because of time constraints or insufficient provisions. Opinions regarding the proximity of green spaces around the airport and port are divergent, with 39% of them viewing them as being in short supply, 38% as being sufficient, and 23% stating that they are none, indicating skewed access and distribution. Participation in organized social activity is about evenly spread, although conceivably positively, 38% had regular take-up, suggesting that where activities are good value and accessible, there can be high levels of take-up (e.g. **Fig. 10**).

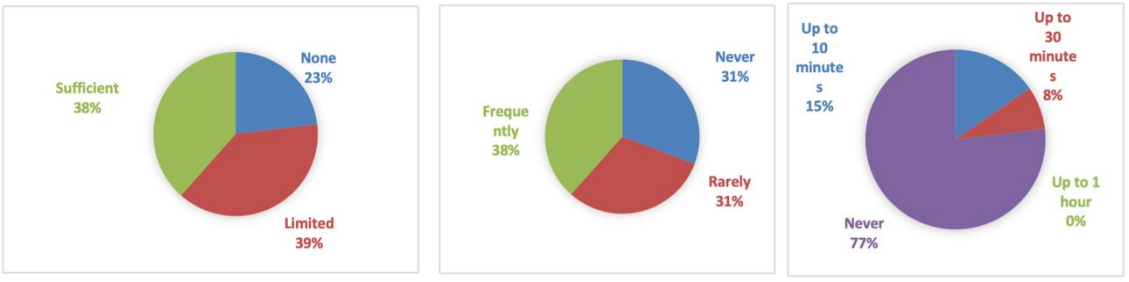


Fig 10. Results of survey analysis - Koper. Graphic: S.Pansinger

Fig 11. Citizens participation in Koper. Photo by: S.Pansinger

In Trieste, focus is laid more strongly on the social and spatial impact of port activities on the city and its citizens. Public opinions indicate strong interest on the part of the public

in improving public space quality, particularly through opening up and urban regeneration of port areas and transition areas (e.g. **Fig. 12**). Over 70% of interviewees reported limited access to public and green spaces within the port area, which they described as having a detrimental effect on quality of life. While the population acknowledges the economic significance of the port, there is a high desire for improved symbiosis between urban activities and the city. Citizens showed a desire for spatial innovations that respect Trieste's historical identity and cultural values, but at the same time integrate new, mixed-use patterns balancing social, ecological, and economic aspects. HubCities digital platform was again defined as a resource that could be employed enhancing massive stakeholder engagement and allowing participatory decision-making (e.g. **Fig. 13**). The ability to collect, share, and interpret jointly local data was found to be crucial in order to make a transformation process clear and well understood by all stakeholders.

A strong 70% of respondents believe there is potential to create or improve public spaces to enhance well-being in their work and community environment (e.g. **Fig. 14**). When asked about ideal features, 37% preferred recreational activities with amenities, another 37% supported general recreation, while 21% favored shopping options and 10% chose green spaces, indicating a preference for active and multifunctional spaces. Additionally, 74% expressed a need to use public space during breaks or free time, highlighting demand for accessible and engaging outdoor environments. Together, these responses point to both the desire and opportunity to design human-centered public spaces that go beyond utility and support daily quality of life.

Fig 12. Position of airport and seaport in Trieste. Graphics: Google Earth

Fig 13. Citizens participation in Trieste. Photo by: S.Pansinger

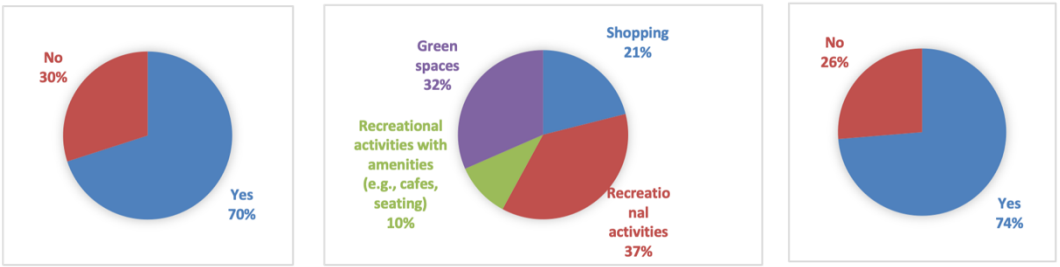


Fig 14. Results of survey analysis - Trieste. Graphic: S.Pansinger

4.1 Cross-Cutting Insights

The cross-site analysis of Graz, Koper, and Trieste reveals that sustainable development of seaport and airport areas is a complex and interconnected issue. Ecological, social, economic, and spatial dimensions are highly interconnected and thus require comprehensive strategies which recognize such complexity. At each of the three case study locations, public participation, specifically through citizen science played an important driver in facilitating various interests and creating public acceptance of the decarbonization processes needed.

Environmental concerns were always ranked a top priority at all places by the citizens. Citizens expressed a prominent requirement of reduced emissions, improved air quality, and noise lowering, which demonstrated a shared environmental awareness and demand for environmental responsibility in planning infrastructure. Besides that, there was a clear and persistent request for open information exchange as well as true participation in planning activities. These findings indicate towards the need for accessible, inclusive modes of participation, facilitated by online platforms such as HubCities.net, making relevant information access available and facilitating collective decision-making.Among the most profiled tensions registered on the sites was one of economic necessity versus ecological sustainability. In Koper, the port's economic function prevailed with emphasis on maintaining it industrial in nature. On the other hand, in Trieste and Graz, public discourse and planning priorities were focused on facilitating

social cohesion and quality of space in dwelling. These are alternative priorities that show the importance of context-sensing planning wherein the local needs and aspirations influence the trends in change strategies.

The third common thread that ran through all three cities was the importance of public and accessible urban space. Citizens demanded more habitable spaces for recreation, cultural life, and social interaction. These were not only essential as an aspect of good urban life, but as protective factors against the impacts of mass-scale infrastructure. Public space thus finds pride of place in social sustainability development and in fostering local attachment to process change. The study shows that online platforms like HubCities are a lot more than they are often portrayed as technical solutions but rather social and communicative drivers. The platform makes valuable planning and environmental data accessible, promotes transparency, and enables broader participation. It unites governance, academia, industry, and civil society stakeholders in one space for collaboration, dialogue, and co-creation. In doing this, it enables the construction of a transdisciplinary community working collaboratively toward sustainable city transformation. The overall findings support the initial objectives of the HubCities project. Participatory planning and Citizen Science combined are not only beneficial but also imperative in addressing the complex ecological, social, and spatial challenges of seaport and airport construction. By using digital technology and collective involvement of citizens, decarbonization processes can be rendered ecologically sound, socially equitable, and spatially balanced. This approach demonstrates that sustainable transformation is not a technical issue but an intensely social and spatial process. HubCities, the model, demonstrates how innovation, participation, and policy come together to promote robust, inclusive, and future-oriented urban futures.

5. Discussion

The three case studies of Graz, Koper, and Trieste yield results that emphasize the complex challenges and transformational opportunities of airport and seaport area sustainable development. Such so-called HubCities, while functioning as vital logistics and economic centers, are also enormous resource consumers and CO_2 emitters. Decarbonization in such an environment can only be successful if overall there is a response to ecological, social, economic, and spatial considerations simultaneously. One of the essential findings in all the sites is the importance of active public participation. The study demonstrates that citizens show great interest and enthusiasm to be involved in designing and implementing sustainable solutions. Citizen Science engagement consolidates the public's awareness of complex interdependencies as well as initiates trust and acceptance. Citizens become active co-producers instead of passive observers, participating significantly in warranting the long-term embedding of decarbonization strategies. Being an airport city, Graz demonstrates how environmental and social efforts

in enhancing quality of urban life in neighboring districts can be successfully implemented. Citizens especially emphasized the need for: more sports and green areas, better air quality, and social spaces that break down barriers between residential and industrial zones. Such interaction of environmental and social factors illustrates the idea of how technical solutions must be compatible with human needs.

The Koper case illustrates a highly sensitive equilibrium among industrial purpose, economic value, and social integration. The port is not only an economic hub but a marker of local identity. Participatory processes managed to establish constructive conversation among the parties involved public administration, industry, and citizens. Sustainability is neither conceived as a solely environmental imperative nor as an environmental imperative in itself but as a way to strengthen social justice and economic resilience. The findings show that decarbonization will be effective only if it is integrated into the local social and economic setting. As a border city, Trieste introduces an additional aspect of refinement with its international link and cross-border collaboration. While environmentally sustainable transformation is encouraged by the populace, they also have genuine concerns regarding competitiveness and jobs. The challenge thus calls for innovative and collaborative solutions that reconcile ecological aspirations with economic stability. In this aspect, collaboration with adjacent regions and cities, assisted through online platforms such as HubCities.net, becomes crucial.

In every location, the HubCities digital platform was a significant facilitator of: knowledge and experience sharing, enhanced transparency, and an engaged citizen participation, including on local borders. together on sustainable solutions, while sharing best practices and success stories

From the perspective of Science and Technology Studies (STS), the use of digital participation tools such as the HubCities platform cannot be regarded as neutral. While the platform facilitated interaction, openness, and transboundary communication, its use also revealed social and digital asymmetries among participants. Better educated citizens with flexible work schedules and reliable internet access were likely to be active, where as industrialand shift workers, particularly those in port settings, had greater limitations on their participation. These dynamics trace out the ways in which infrastructures of digitality, although designed to democratize planning, are prone to reproduce existing class-based, labor-divided and technological-accessed inequalities. The variation between the three cities also reflects their different economic and social profiles. Graz's service and research-oriented economy allowed for more widespread and diverse civic engagement. Conversely, Koper's industrial-strength and Trieste's blended logistic and heritage setting produced more selective and issue focused engagement. These variations underscore that sustainable transformation citizen involvement is not merely a question of motivation but also of social status, economic livelihood and local institutional habit. On a broader level, the findings resonate with conventional STS arguments about the non-neutrality of technological systems. Seaports

and airports are not merely transport or logistics infrastructure but socio-technical systems that embody political agendas and economic hierarchies. Within such globally networked and large-scale contexts, local participatory processes such as those stimulated through HubCities can make open discussion, increase transparency, and disrupt planning trends, but must be constrained by the very character of acting upon global economic systems. Regardless, these participatory spaces are critical spaces of negotiation, in which collective awareness and small-scale interventions can collectively reorder institutional practices towards more sustainable and equitable possibilities. Despite these positive outcomes, there are still challenges to be addressed, assuring long-term participation procedures, security of the data, and electronic access, as well as the construction of political and bureaucratic institutions that institutionalize participatory methods and ensure their effectiveness. This research provides significant findings of relevance to many other urban regions confronting similar sustainability issues. It offers a vision of the possibilities of technological innovation, participatory planning, and social mobilization in facilitating socially equitable and sustainable urban change. HubCities demonstrates how technological systems are socio-technical articulations with definite interests and power relations. Through the demonstration of such underlying dynamics, the project relocates decarbonization away from being merely a technical issue but as one of social and spatial negotiation whereby sustainability is co-produced in the course of relations among citizens, planners, and institutions. HubCities can therefore be a driver in achieving climate goals while concurrently building sustainable, inclusive, and resilient cities for their citizens.

6. Conclusion

The present research being conducted as part of the HubCities project represents a valuable contribution to the research and utilization of sustainable development strategies for energy-intensive urban centers — i.e., airport and seaport areas. The entirety perspective, in turn, grounded on unifying the ecological, social, economic, and spatial elements of decarbonization, points to the multifaceted nature as well as the necessity of multidisciplinary examination of such urban regions. First and foremost, the empirical evidence confirms that effective decarbonization of HubCities can be achieved only through active involvement by local communities and suitable stakeholders in the future as well. Not only does citizen science present itself as a participatory tool, but also as a methodological interface between science, planning, and society. Public engagement has an extremely important role in guaranteeing increased transparency, acceptance of climate protection measures, and urban change in a socially equitable manner. This is particularly important in view of the large size of airport and port infrastructures and multi-dimensional stakeholder interests involved. Second, the case studies of Graz, Koper, and Trieste highlight the need for a strategy that builds on an

integrated approach that goes beyond narrowly technical or economic thinking. The close interrelation between social quality of life, ecological compatibility, and economic competitiveness becomes all the more important. The results reveal that reduction measures and efficiency measures for emissions are only effective in the long term when complemented with socio-spatial strategies enhancing residents' well-being as well as inclusive social engagement. Multifunctional, accessible, and high-quality public spaces developed in HubCities lead to social cohesion and enhance public support for necessary infrastructural and technical change. Third, the study shows that intermunicipal cooperation and spatial connectedness are preconditions for success in decarbonization. Intra-border cooperation like in the case of Trieste puts more pressure on government and coordination systems, but also opens new innovative possibilities. Websites like www.hubcities.net have an important role to play by promoting knowledge transfer, facilitating participatory processes, and building a transnational community of practice. The systematic use of such tools makes it possible to collect experiences, good practices, and knowledge to be accumulated beyond territorial boundaries and adaptively incorporated into decision-making. Fourth, the project emphasizes the need to strengthen institutional infrastructures to the sustainable development of HubCities. This entails: Developing legally binding processes of participation, Institutionalizing citizen science, and Designing flexible and adaptive planning tools that respond to multidimensional requirements. Such systems require revamping planning and administration — towards collective and learning-oriented systems of administration capable of handling complexity and ensuring long-term flexibilities. In conclusion, HubCities provide a versatile model of integrating decarbonization, social integration, and spatial sustainability in energy-dense urban systems. The interplay between participatory science, technological innovation, and social integration is a model which can be replicated in other urban networks. The project, thus, not only contributes to addressing climate targets but also increases societal resilience and promotes future-proof urban environments. This multi-perspective approach will become increasingly important in the years ahead, as cities confront growing challenges from climate change, urbanization, and social inequality. Further explore the interplay between technological innovation, participatory processes, and institutional frameworks, and Assess the scalability and transferability of the developed concepts to other urban infrastructures and regions. It is only by taking such an integrative and adaptive strategy that the long-term security of the sustainable transformation of HubCities can be guaranteed and made a model for other energy-consuming urban agglomerations.

7. References

Ažman Momirski, L. (2004) ‘The Port of Koper: The Youngest Modern North Adriatic Port’, Portus, 7, pp. 70–75.

Ažman Momirski, L. (2021) ‘The Resilience of the Port Cities of Trieste, Rijeka, and Koper’, Journal of Urban History, 47(2), pp. 293–316. doi:10.1177/0096144220926600.

Bonney, R., Shirk, J.L., Phillips, T.B., Wiggins, A., Ballard, H.L., Miller-Rushing, A.J. and Parrish, J.K. (2014) ‘Next steps for citizen science’, Science, 343(6178), pp. 1436–1437.

Cipriani, L. (2014) Ecological Airport Urbanism: Airports and Landscapes in the Italian NorthEast. Trento: Aracne Editrice.

Forester, J.F. (1999) The Deliberative Practitioner: Encouraging Participatory Planning Processes. Cambridge, MA: MIT Press.

Fusco Girard, L. (2013) ‘Toward a Smart Sustainable Development of Port Cities/Areas: The Role of the ‘Historic Urban Landscape’ Approach’, Sustainability, 5, pp. 4329–4348. doi:10.3390/su5104329.

Graham, S. and Marvin, S. (2001) Splintering Urbanism: Networked Infrastructures, Technological Mobilities and the Urban Condition. London/New York: Routledge.

Hecker, S., Haklay, M., Bowser, A., Makuch, Z., Vogel, J. and Bonn, A. (eds.) (2018) Citizen Science: Innovation in Open Science, Society and Policy. London: UCL Press.

Healey, P. (2006) Urban Complexity and Spatial Strategies: Towards a Relational Planning for Our Times. London: Routledge.

Hilleri, L. and Sieverts, T. (2004) Zwischenstadt: Ein neues Leitbild für die Urbanisierung Europas. München: Vieweg.

Konvitz, J. (2012) ‘Contemporary Urban History: What the Study of Port Cities Implies for Evidence, Methodology, and Conceptualization’, Journal of Urban History, 39(4), pp. 801–806. doi:10.1177/0096144212470248.

Maček, M. (2016) Novo letališče obalne regije: idejna zasnova letališča v Mestni občini Koper (Diplomsko delo). Fakulteta za arhitekturo, Univerza v Ljubljani.

Morgan, D.L. and Bottorff, J.L. (2010) ‘Advancing Our Craft: Focus Group Methods and Practice’, Qualitative Health Research, 20(5), pp. 579–581. doi:10.1177/1049732310364625.

Nielsen-Bohlman, L., Panzer, A.M., Kindig, D.A. and Institute of Medicine (eds.) (2004) Health Literacy: A Prescription to End Confusion. Washington, D.C.: National Academies Press.

Norberg-Schulz, C. (1979) Genius Loci: Towards a Phenomenology of Architecture. New York: Rizzoli.

Pansinger, S. (2017) 'Gestalt Sustainability – the future field of action to reduce our ecological footprint', Der Standard. Available at: https://www.derstandard.at/story/2000058692472/wie-man-nicht-orte-in-orte-fuer-menschen-verwandelt

Pansinger, S. (n.d.) 'HubCities - A New Approach to Sustainable Development of Airport and Seaport Territories through Citizen Science'. Available at: www.hubcities.net

Pansinger, S. and Förster, J. (2018) 'Airport neighbourhoods as future regional development areas for resource awareness and gestalt sustainability: SmartAirea', WIT Transactions on Ecology and the Environment, 2. Available at: https://bit.ly/2PyUuaU

Pansinger, S. and Prettenthaler, F. (2023) 'Gestalt Sustainability', disP - The Planning Review, 59. doi:10.1080/02513625.2023.2229626

Pansinger, S. and Ažman Momirski, L. (2022) 'Air- | Seaport cities: on metropolitan territory of HubCities', REAL CORP 2022: Mobility, Knowledge and Innovation Hubs in Urban and Regional Development.

Ports (2017) European Commission: Maritime – Ports. Available at: https://transport.ec.europa.eu/transport-modes/maritime/ports_en

Transport EU (2022) European Commission: Air – Airports. Available at: https://transport.ec.europa.eu/transport-modes/air/airports_en

Vazquez, E.F. and Morollon, F.R. (2012) Defining the Spatial Scale in Modern Regional Analysis: New Challenges from Data at Local Level. Springer.

Continuing Education in HTA for Digital Health Integration

Mie Basballe Jensen, Tom Børsen, Frederik Albert Berthing, Lucas Klingenberg Mathisen, Olivia Bjørnholdt Overgaard, Sisse Rej Rasmussen, Sasha Sofie Mie Rasmussen, Christian Ditlev Zinck

Aalborg University, Denmark

DOI 10.3217/978-3-99161-062-5-012, CC BY 4.0

Abstract. This paper explores how continuing education can support more context-sensitive and ethically grounded enactments of digital health technologies in the Danish healthcare system. Based on 20 semi-structured interviews with health actors across clinical, educational, managerial, innovation and policy domains, we analyse how different professionals perceive current enactments of health technologies and the role of continuing education. Our analysis reveals widespread concern over time scarcity, fragmented responsibilities, and lack of shared vocabularies across professional domains. Interviewees call not only for technical training, but for educational spaces that support critical reflection, ethical awareness, and cross-professional dialogue. In response, we present Health Technology Assessment 2.0 (HTA 2.0), a framework developed for continuing education. Drawing on both inductive and deductive coding, we examine how its six dimensions (Technology, Economy, Environment, Organisation, Patient/Citizen, and Ethics) resonate with everyday practice and healthcare actors' concerns. We suggest the potential of HTA 2.0 to act as a boundary object: structuring shared reflections while accommodating different professional viewpoints. We conclude that continuing education should not aim for consensus but provide structured arenas where health actors can explore challenges, reflect on dilemmas, and co-develop meaningful approaches to digital transformation.

1 Introduction

1.1 Demographic Changes and the Growing Healthcare Demand

The Danish healthcare system is increasingly shaped by demographic and structural changes. An aging population, the rising prevalence of chronic diseases, and growing citizen expectations are converging to place considerable demands on both healthcare services and the healthcare professionals delivering them (Højgaard & Kjellberg, 2017).

These pressures are not merely numerical; they challenge the core organization and sustainability of care.

By 2036, the number of citizens over the age of 80 is expected to have nearly doubled compared to 2016, while the working-age population continues to decline (Højgaard & Kjellberg, 2017; Hansen et al., 2022). This demographic 'double pressure' implies that more people will require complex care, but fewer will be available to deliver it. Compounding this, the demand for healthcare professionals is rising significantly, with projections estimating a need for 44,000 additional employees in the public sector by 2030 just to maintain current service levels (KL, 2022). Yet recruitment and retention remain major challenges: the number of vacant nursing positions has increased, and resignation rates among healthcare staff have surged by 50% from 2020 to 2024 (Sundhedsmonitor, 2024).

These developments are mirrored in the increasing complexity of care. Patients with more than one chronic condition require longer, cross-sectoral treatment trajectories, demanding strong coordination, new types of competencies, and flexible systems. For instance, patients with three or more chronic conditions generate healthcare costs up to eleven times higher than those without any (Højgaard & Kjellberg, 2017).

1.2 Technology as a Proposed Solution?

In response to mounting structural challenges, Danish healthcare policy has increasingly turned to digital health technologies as a potential solution (Indenrigs- og Sundhedsministeriet, 2023). Strategies such as the National Strategy for Digital Health (Sundheds- og Ældreministeriet, KL & Danske Regioner, 2018), promote an integrated, citizen-centered healthcare system supported by scalable, interoperable digital solutions. *Local Government Denmark* (KL, 2022) and *Danish Regions* (Danske Regioner, 2022) highlight technologies, such as medication robots and video consultations, as pragmatic tools to ease workloads, enhance patient autonomy, and improve efficiency.

The Danish Resilience Commission (Robusthedskommissionen, 2023) emphasizes technology's role in addressing staff shortages by automating tasks, enhancing patient self-care, and supporting differentiated service models. The Commission recommends structural reforms to accelerate adoption, including modernized regulation and funding mechanisms. Yet stakeholders also caution against simplistic 'technological quick fixes' (Langstrup & Gjødsbøl, 2023) and point to major implementation challenges: insufficient governance, inconsistent evidence assessment, lack of guidance on how to implement and use health technologies in practice, as well as limited continuing education.

These strategies stress shared infrastructure and digital standards, encouraging locally driven innovation to be scaled nationally. However, the success of such innovation relies heavily on implementation capacity, professional engagement, and systematic knowledge sharing. Digital technologies may offer great potential, but successful

integration is contingent on meaningful implementation, local anchoring, and proper workforce training (Ugeskriftet, 2018a, Ugeskriftet 2018b). Health professionals often experience frustration and encounter challenges when new tools are introduced without time, support, or adaptation of workflows (Jensen & Børsen, 2024).

At the EU level, *Regulation 2021/2282* reflects a growing recognition that the assessment and implementation of health technologies is a complex process. This regulation aims at harmonising the approach to *Health Technology Assessment* (HTA) across member states and reducing fragmentation and duplication in assessment procedures. It was introduced to meet the challenges developers face when navigating multiple, parallel national requirements that is said to delay innovation and increase costs. By establishing joint clinical assessment procedures to support national decision-making, the EU seeks to streamline evidence processes while respecting local healthcare contexts (European Parliament and Council, 2021).

The *WHO Regional Digital Health Action Plan for the European Region 2023–2030* (World Health Organization, 2022), also identifies digital transformation as a key accelerator of resilient and people-centered healthcare systems. It outlines strategic priorities including governance, literacy, evidence-building, and equity in digital health adoption. It is emphasized that digital innovation must be driven by real-world health needs, respect professional expertise, and empower citizens.

1.3 A gap Between Technological Potential and Real-World Use

While digital health technologies hold immense promise for improving healthcare efficiency, quality, and access, real-world use often falls short of this potential. The implementation of new digital technologies tends to be far more complex than anticipated, especially when introduced into already strained healthcare systems. Technologies that appear beneficial on paper frequently lead to unintended consequences such as increased workload, fragmented workflows, and staff frustration. An example is the implementation of the Epic based electronic health record implemented in two Danish regions in 2016. The rollout was followed by major workflow disruptions, data integration failures, and sharp increases in time spent on clinical documentation. Reports describe patient injuries linked to system errors, and five years after go-live, one third of users still express dissatisfaction. The case illustrates how large-scale digitalization can compromise care quality and staff wellbeing when technological ambition outpaces organizational readiness (Hertzum, Ellingsen & Cajander, 2022).

A large-scale analysis has revealed that many digital interventions fail to reduce staff time or improve productivity. Among 467 reviewed studies, over 30% showed no or even negative impact on healthcare staff time (Shemesh et al., 2025). The reasons were primarily linked to poor usability, lack of training, additional administrative burdens, and

failure to adapt existing workflows. The findings challenge the widespread assumption that procurement of digital tools alone is enough to generate meaningful benefits.

Frontline experiences often illustrate a stark mismatch between policy ambitions and everyday realities. A study of public sector digitalization describes how healthcare professionals must continuously adapt to shifting digital systems while juggling core responsibilities. The result is a workday marked by system fragmentation, constant change, and limited time for actual patient-centered tasks. As Oskarsen and Bratteteig (2024) underline, the additional time and resources required for technology implementation often constitute invisible work that remains poorly recognized.

These experiences also reflect a broader structural challenge: the organizational context is rarely ready to absorb the full impact of digital change. Agile development practices may enable rapid software iteration, but they often fail to align with the slower, highly interdependent nature of clinical work. As a result, system updates and new functionalities can outpace organizational capacity for adaptation, creating continuous disruption and frustration among staff (Oskarsen & Bratteteig, 2024).

In parallel, HTA rarely documents clinician time as a key metric. This is documented in a recent scoping review of telemedicine trials. Among the 78 included studies, only four measured clinician time directly, and most found no significant difference between telemedicine and standard care (Kidholm et al., 2024). Despite this, time use is rarely included as a key evaluation parameter in formal HTAs. As a result, current assessment practices risk overlooking one of the most pressing challenges facing healthcare systems today i.e., the shortage of time and personnel.

The result is a situation where digital technologies are introduced with high expectations, but limited awareness of the conditions necessary for successful integration. Without robust implementation strategies, local adaptation, and investment in staff training and engagement, the benefits of digitalization risk remaining theoretical.

1.4 Bridging the Gap through Continuing Education

Frameworks increasingly recognize that digital transformation requires not only smarter technologies, but smarter learning systems. *The Danish Ministry of Higher Education and Science* (Uddannelses- og Forskningsministeriet, 2023) identifies continuing education as a strategic lever for supporting innovation in the life sciences and healthcare sectors. Particular emphasis is placed on interdisciplinary, practice-oriented formats that support professionals in dealing with both technological and societal challenges. Yet, the existing continuing education landscape remains fragmented. A national analysis by the *Danish Center for Social Science Research* (VIVE, 2023) shows that while non-formal digital and clinical training opportunities exist, they are often scattered, short-term, and poorly coordinated making them difficult to navigate for time-constrained healthcare professionals. In response, the *Danish Ministry of Higher Education and Science*

launched a targeted funding scheme under the national Life Science Strategy to support the development of continuing education initiatives that address digitalisation, automation, and technological change in healthcare (Uddannelses- og Forskningsministeriet, 2023). This project is funded by that programme and reflects a broader political recognition of digital transformation not only requires new tools, but also new professional competencies.

To conceptualize the potential of continuing education in bridging 'the implementation gap' in digital healthcare, we draw on perspectives from *Responsible Research and Innovation* (RRI) and *Science and Technology Studies* (STS). RRI calls for embedding anticipation, inclusion, reflexivity and responsiveness into the development and governance of technologies (Stilgoe et al., 2013). In healthcare, this entails recognizing that technologies do not only solve problems they also reconfigure professional identities, ethical obligations, and the distribution of work. Rather than prescribing fixed solutions, contemporary RRI approaches emphasise contextual translation and value negotiation in local settings (Boenink & Kudina, 2020). This is the type of work that continuing education could potentially support, by creating structured spaces where professionals can explore emerging dilemmas, voice concerns, and develop anticipatory competences before innovations become entrenched.

STS complements this view by offering conceptual tools to understand how reflection and collaboration happen in practice. One such concept is critical proximity (Amanatidis & Børsen, 2024) that refers to the ability to stay close enough to practice grasping its constraints, while maintaining enough distance to critically engage with routines and institutional logics. Continuing education may offer a site for cultivating this stance allowing professionals to examine real-world dilemmas without the pressure of immediate decision-making, fostering a mode of inquiry that is both grounded in practice and reflexive.

STS can also open for an understanding of how cross-professional collaboration around HTA can unfold despite different views, through the concept of boundary objects (Star & Griesemer, 1989). Boundary objects are concepts, frameworks, or artefacts that are flexible enough to adapt to local needs while retaining a stable identity across domains. In the context of this project, HTA continuing education can be seen as a possible boundary object (Levina & Vaast, 2005). We will use this concept to explore if continuing education in HTA can create a common ground where health actors, with different roles and perceptions, can engage in constructive dialogue about assessment and implementation of digital health technologies.

Thus, in this paper, we approach continuing education not as a vehicle for teaching professionals how to use technologies, but as an arena for collective reflection and dialogue about how technologies shape care, professional judgement, and organizational practice. Rather than focusing on operational proficiency, we conceptualize continuing

education as a space for enactment, where professionals actively examine, discuss, and negotiate the meanings, risks, and opportunities of digital health tools in their own contexts. This shift from *use* to *enactment* highlights education as a reflective and anticipatory practice rather than a purely technical one.

Taken together, these frameworks point to continuing education as a promising but underutilized arena for fostering more reflexive and context-sensitive digital innovation. It is not a silver bullet, but it may offer an important entry point for engaging with ethical, organizational, and practical tensions that often complicate implementation of health technologies. Thus, our research question addressed in this paper is:

> *How do healthcare actors perceive continuing education in Health Technology Assessment as a means to bridge the gap between the potential of digital health technologies and their everyday enactment in clinical practice?*

2. Method

To answer this research question, this study employes a qualitative and interview-based research design that explores how healthcare actors understand and assess digital health technologies in practice, as well as how they see and desire continuing education. Semi-structured interviews were chosen as they allow for both consistency and flexibility across conversations, enabling participants to reflect on concrete experiences while also articulating broader concerns and priorities. This approach can capture nuanced, context-dependent insights needed to inform the development of continuing education frameworks.

In line with the exploratory aim of the study, the analysis followed a two-step design combining inductive and deductive coding. In the first, inductive phase, emergent themes were identified from the interview material to capture how health actors describe everyday experiences, tensions, and needs related to digital technologies. In the second, deductive phase, these insights were revisited through the analytical lens of the *Health Technology Assessment* 2.0 (HTA 2.0) framework (further explained in Section 3.2) to explore how informants' reflections related to its six dimensions: Technology, Economy, Organisation, Patient/Citizen, Ethics, and Environment. HTA 2.0 was selected because it provided a structured yet flexible framework for identifying which aspects of technological change participants emphasized, neglected, or contested. Rather than evaluating technologies themselves, the framework was used to map how different aspects and implications surfaced in the informants' reflections.

2.1 Semi-Structured Interviews

The data collection consisted of 20 semi-structured interviews with actors across the Danish healthcare sector. Participants were selected via purposive sampling to ensure diversity in professional backgrounds, institutional affiliations, and hands-on experience with digital health technologies. Interviewees included clinical staff, educators, policy advisors, innovation consultants, and representatives from hospitals, municipal services, and professional organizations. To provide a structured overview of participant diversity, the 20 informants were grouped into four main categories based on their professional affiliation and role: (●) Innovation and digitalization units, (●) Healthcare professionals, (●) Academics and educators of Healthcare Professionals, and (●) Professional and regulatory organizations in healthcare in the healthcare sector. *Figure 1* illustrates the distribution of interviewees across these categories.

Color	Group	Informants
●	Innovation and digitalization	2, 3, 6, 7, 9, 10, 11, 17
●	Healthcare professionals	12, 13, 16, 20
●	Academics and Educators of Healthcare Professionals	4, 5, 14, 15, 18,
●	Professional and regulatory organizations in healthcare	1, 8, 19

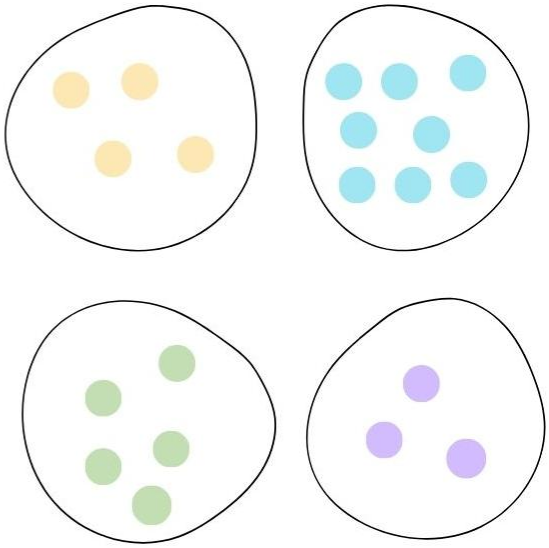


Figure 1: Overview and grouping of interviewees

To guide the interviews, a semi-structured interview guide was developed and used. The guide covered a range of themes: from procurement and assessment of digital technologies to implementation strategies, organizational involvement, sustainability considerations, and the perceived need for continuing education to bridge the implementation gap. Questions were tailored to elicit both evaluative and experiential insights for instance, how needs for new technologies are identified, how success is defined during implementation, and what competencies are seen as lacking or essential for engaging with digital health tools.

The interviews were conducted online, lasted approximately 60 minutes, and were audio-recorded with the participants' informed consent. Transcriptions and minutes of the

interviews were prepared and anonymized. Anonymization was carefully negotiated to protect individual identities while still retaining relevant information about participants' institutional and professional contexts. Each participant was consulted on how they are presented in the final output.

The interviews were analysed using the hermeneutic circle, an iterative method that moves between parts and wholes to refine understanding continuously. This approach allowed us to identify and relate individual perspectives to broader institutional and sectoral patterns, resulting in a rich understanding of how digital health technologies are assessed, negotiated, and made to work in practice.

Through the analysis we identify illustrative excerpts to be used in this paper. To validate the interpretations and ensure accurate representation, the selected quotes and their contextual framings were shared with the interviewees, who were given the opportunity to revise, clarify, nuance, or retract their contributions. This feedback loop strengthened the trustworthiness of the material and ensured that the analysis accurately reflected the intentions and insights of the participants.

3 Findings and Analysis

The following section presents the empirical findings based on 20 semi-structured interviews with healthcare actors. The section is structured in three parts: first, we outline shared and divergent perspectives among healthcare actors (3.1); second, we present five inductive themes that illustrate practical tensions and knowledge needs (3.2); and finally, we apply the HTA 2.0 framework deductively to examine how these issues align with its six analytical dimensions (3.3).

3.1 Shared and Divergent Health Actors Perspectives on Continuing Education

The interview material reveals both shared concerns and meaningful divergences in how different healthcare actors perceive digital transformation and the role of continuing education. Across the four main groups (innovation and digitalization consultants, healthcare professionals, educators and academics, and representatives of professional and regulatory organizations) there is strong agreement that digitalization cannot succeed through technical training alone. Informants across roles and professions emphasize the need for competencies that include critical reflection, contextual understanding, and ethical awareness.

More informants stress that continuing education is essential not only to enable safe and effective use of digital tools, but also to support professional judgement and maintain critical and contextual reflection. As one interviewee put it:

> *'We still need to keep our critical gaze, and that makes it even more important to develop some kind of competence, so we don't end up causing too many unintended incidents'* (I5, Academics and Educators).

Another reflected on the strain continuing education might place on healthcare professionals, noting that:

> *'It has become part of our work life that we must continue educating ourselves in technology while we work. And I think that's a cruelly underappreciated part of being a healthcare professional. The job of a healthcare worker is to deal with illness and human life. When we are asked to learn something new, we risk making mistakes, both for ourselves and others*.' (I11, Innovation and digitalization).

Significant challenges were raised regarding the feasibility of current and potential future educational offerings. Time scarcity, organizational pressure, and insufficient managerial support were cited as significant barriers. In this context, continuing education is not merely a technical fix but a site of negotiation about institutional priorities, the conditions under which professionals can learn, and the values that guide digital transformation. These insights align with recent national policy frameworks, including those from the *Danish Ministry of Higher Education and Science* (Uddannelses- og Forskningsministeriet, 2023) and the *Danish Centre for Social Science Research* (VIVE, 2023) analysis, which call for more coherent, practice-oriented educational strategies.

Despite this consensus, significant differences exist regarding how continuing education should be structured and what it could and should achieve. Health professionals emphasize safeguarding care quality, relational work, and patient safety. They call for educational formats that respect resource constraints and support decision-making under pressure. In contrast, innovation and digitalization consultants often frame continuing education as a lever for accelerating implementation, aligning practices with strategic goals, and improving efficiency. Educators stress the importance of flexible, practice-based learning grounded in pedagogical principles, while representatives of professional organizations highlight structural issues such as fragmented training opportunities, lack of coordination, and the need for clearer governance.

These differences reflect not just distinct roles, but different institutional logics: managerial, clinical, pedagogical, and policy driven. Each logic informs specific ideas about what counts as valuable knowledge, acceptable risk, and legitimate implementation. The result is not only misalignment of expectations but also practical tensions in how educational initiatives are understood and prioritized. These converging and diverging perspectives highlight the need for continuing education that is both flexible and dialogical, enabling professionals to navigate multiple logics and reflect across roles.

In the following section, we turn to inductive themes that further illuminate how these tensions and needs play out in everyday clinical and organizational settings.

3.2 Inductive Themes Emerging from the Interviews

Our inductive analysis generated five cross-cutting themes that illuminate the practical tensions and point to perceived knowledge needs faced by healthcare professionals engaging with digital technologies.

3.2.1 Impact of Technology in Practice

Interviewees highlight the dual role of technology as both enabler and obstacle. While acknowledging its potential, they frequently cite frustrations with poor usability, system fragmentation, and mismatch with clinical workflows.

> *'The dream is to have technology either optimize/streamline specific types of work, so you can avoid doing manual tasks, for example, and instead shift time toward performing the core task.'* (I11, Innovation and digitalization)

This statement captures the desire for technology to free up time and resources for healthcare professionals, enabling them to focus more on their primary responsibilities. However, it also hints at the practical challenges in achieving this ideal. The theme also reveals a pragmatic knowledge boundary, where management's drive for efficiency contrasts with clinicians' emphasis on professional judgement and relational care. This divergence illustrates how the same technology can be valued differently depending on institutional priorities, requiring spaces of negotiation to align its intended and experienced effects.

3.2.2 Structural Barriers

Informants identify systemic issues such as insufficient training, lack of time, and siloed decision-making. These obstacles make it difficult to translate strategic goals into practice.

> *'There is a lack of someone with the ongoing responsibility to maintain and possess the necessary competencies when it comes to technology implementation in healthcare.'* (I9, Innovation and digitalization)

This highlights how the absence of continuous responsibility for competency development contributes to the challenges in sustaining effective technology implementation. More of the interviewees mention structural fragmentation amplifies syntactic boundaries: the absence of a shared vocabulary between technology developers and clinical staff complicates decisions about what constitutes 'solid evidence'. For example, expectations rooted in evidence-based medicine often clash with the more situated, practical impacts of digital tools in practice.

3.2.3 Critical Tech Literacy of Healthcare Professionals

More informants underline that the problem is that healthcare professionals are underprepared to engage with technology beyond operational use. They emphasize the need for tools and training that support critical thinking and ethical reflection regarding technology in healthcare. As one informant explained:

> *'There is a difference between knowing how to navigate social media at home and having a professional technological literacy, because you have to communicate differently with patients, for example when chatting with them.'* (I5, Academics and Educators)

This distinction underscores the importance of developing profession-specific digital competencies that go beyond everyday technology use, enabling healthcare professionals to communicate effectively and ethically in clinical contexts. These findings point to semantic boundaries concerning the concept of 'good technology'. While decision-makers may equate it with cost-efficiency or scalability, practitioners emphasise alignment with professional values, user needs, and patient safety. This divergence illustrates tensions between different institutional logics (Riiskjær, 2014). Healthcare is a pluralistic field, where professional, managerial, and market-oriented logics coexist and often clash. These frictions shape how new technologies are interpreted, resisted, or adapted in practice.

3.2.4 Leadership and Organisational Change

Stronger leadership engagement and clearer implementation strategies are consistently identified as crucial for successful technology adoption. One informant highlights the importance of making technology-related tasks a core responsibility within performance assesments:

> *'If it is a core task you are measured on, then it also becomes a managerial focus. Because we act in accordance with how we are measured, that's how we are structured.'* (I10, Innovation and Digitalisation)

This statement emphasizes the necessity for leadership to prioritize and systematically support change management, as accountability drives organizational focus and action. The theme also reinforces the importance of addressing pragmatic boundaries at an organisational level. Leadership structures often overlook the continuous competence development required to absorb technological change, thereby rendering this work 'invisible' in formal systems of recognition and assessment.

3.2.5 Contextualization and Integration of Technology

Healthcare actors hold diverse interpretations and attitudes toward how technological solutions should be implemented. This diversity is illustrated by an informant who notes the importance of bridging different professional perspectives: *'They are two different worlds, and it is really important that they meet in the same room.'* (I2, Innovation and digitalization)

This quote highlights the challenges caused by a lack of shared language and differing realities between IT professionals and clinicians, which can hinder collaboration. The informant further emphasizes the need for open dialogue about needs and challenges to ensure solutions truly address user requirements and to tailor technologies to local contexts to balance scalability with practical usability.

Across all five themes, the interviews reflect a landscape marked by overlapping knowledge boundaries. These boundaries reflect divergent institutional aims, competing definitions of value, and incompatible evidence standards, and do not only hinder collaboration but also shape what kinds of technologies are adopted, resisted, or adapted. Continuing education has the potential to act as a boundary infrastructure that makes these frictions visible and negotiable through shared inquiry and reflective dialogue.

3.3 From Emergent Needs to Structured Assessment: Rethinking HTA

The five inductive themes outlined above reveal complex tensions, unmet needs, and interpretive boundaries in the everyday use of digital technologies in healthcare. To further examine how these challenges map onto existing assessment frameworks, we applied a deductive coding strategy based on an expanded version of the classical HTA framework. In the initial deductive phase, interview responses were coded according to the traditional HTA domains: Technology, Economy, Organisation, and Patient/Citizen (Børsen, 2025). However, it quickly became clear that additional critical aspects, especially ethical dilemmas, and environmental impacts, are consistently emphasised by participants but absent in the formal framework.

This observation aligns with trends in international health policy developments. The *EU Regulation 2021/2282* calls for a harmonised and evidence-based approach to HTA across Europe, while simultaneously underlining the need to implement technologies in ways that are efficient and sustainable for both patients and society (European Parliament and Council, 2021). Similarly, the Organisation for Economic Co-operation and Development(OECD) has introduced the concept of anticipatory governance, which advocates for early reflection on ethical and environmental concerns in the development and assessment of new technologies (Robinson et al., 2023). These broader developments directly support interviewees' calls for more practical and reflective

models: *'If we had an assessment approach that included ethics and sustainability locally at hospitals, I truly believe it would have an impact.'* (I1, Professional and regulatory organizations)

Thus, we introduce the HTA 2.0 framework (Jensen & Børsen, 2025) that is visualized in Figure 2. This updated model retains the four classical domains but adds two essential dimensions: Ethics and Environment. These are not included to increase complexity, but to surface issues that are already part of practitioners' everyday experience yet often remain invisible in formal assessments.

Figure 2: The HTA 2.0 framework (Jensen and Børsen, 2025). The model expands traditional HTA with ethical and environmental dimensions to support reflective decision-making in clinical settings. The star-shaped layout visually illustrates how the six dimensions are interlinked, indicating that they do not operate in isolation but continuously influence and shape one another in practice.

HTA 2.0 should not be understood as a complete model, but as a starting point for developing a more practice-oriented approach to health technology assessment. The six dimensions form an initial analytical structure that can be further specified through concrete methods and reflective questions in educational settings. For instance, ethical aspects may be explored through deliberative and ethical inquiry methods, technological aspects through user testing, and organisational aspects through mapping exercises of decision and responsibility chains as well as through observations of workflow integration. In the context of continuing education, these methodological elements will be developed iteratively together with participants, allowing the model to evolve as both an analytical and pedagogical tool. The intention is to further develop the model throughout

the project, as we consider it a dynamic rather than a static framework. In this sense, HTA 2.0 functions as a boundary object flexible enough to invite interdisciplinary dialogue, yet stable enough to provide a shared language for exploring how digital technologies transform healthcare practices.

The HTA 2.0 framework thus serves a dual purpose: it functions both as an analytical and a pedagogical tool to support reflection, learning, and informed decision-making. In the following section, we apply its six dimensions deductively to the interview data to explore how healthcare actors articulate challenges and priorities related to each domain.

3.3.1 Technology: Assessment, Integration, and Professional Agency

The interviews reveal that digital technologies hold significant promise, but their actual value is difficult to document systematically. Several respondents express concern over the lack of usable evidence demonstrating real-world effects, particularly in terms of workflow improvements and labour savings. One regulatory informant questioned the asymmetry between pharmaceutical and technological approval practices: *'No one today would take a pill if it hasn't been tested, so why would we implant or monitor with technology if we don't know it works?'* (I1, Professional and regulatory organization).

This highlights a widespread concern that technologies are often introduced before robust clinical documentation is available. Technological fatigue emerges as a barrier, especially when early implementation is poorly anchored, or communication is lacking. As one healthcare professional explained, even those tasked with promoting new systems can lose motivation when scepticism dominates:

> *'It is incredibly difficult to act as a super-user when so many people have already formed a negative opinion about the technology. I found it very hard to be the front person for something like that'* (I12, Healthcare professionals).

Technologies that are not aligned with existing work routines are experienced as burdensome rather than supportive. This underlines a shared call across groups for critical engagement and professional ownership in assessing whether technologies truly meet clinical and organizational needs.

Innovation and digitalization actors specifically describe technology as a strategic enabler of system-wide transformation, focused on scalability, interoperability, and structural efficiency. Yet many also point to a gap between ambition and execution. As one consultant noted: *'You can't implement change by dropping a new system in people's inbox. There needs to be dialogue and planning'* (I2, Innovation and digitalization). Another added: *'There is a lack of people with a clear mandate to secure implementation and follow-up'* (I9, Innovation and digitalization).

These reflections show an awareness that success depends not only on technical solutions, but on leadership, and institutional support.

Healthcare professionals approach technology from a pragmatic standpoint, grounded in clinical workflow and patient care. Several describe frustration with tools that increase documentation without improving efficiency: *'It doesn't help me finish my shift faster, it just adds more clicks'* (I16, Healthcare professional). Their experiences point to a recurrent usability gap and to the risk of alienation when new tools are introduced without sufficient adaptation or consultation.

Educators and academics view technology through a pedagogical and epistemic lens. Their emphasis lies not only on operational skills, but also on fostering critical reflection on how digital tools shape professional judgment, relations, and responsibilities. As one educator explained, *'It's about developing a professional understanding of technology, not just using it, but reflecting on what it does to the practice'* (I14, Academics and educators). Another highlighted the importance of making this reflection an integral part of learning: *'We don't teach technology as a separate thing, it's integrated across subjects because it's part of the profession'* (I15, Academics and educators).

Professional and regulatory actors emphasize safety, accountability, and system-level coherence. They express concern over the lack of systematic, transparent assessment processes at the local level and warn against premature adoption (I1, Professional and regulatory organization). Concerns were also raised about data ownership and the risks of dependency on commercial platforms.

Taken together, these perspectives reveal that 'technology' is not a neutral artefact, but multifaceted. While innovation actors focus on systemic impact, clinicians stress usability, educators promote reflective learning, and regulators demand robust assessment. HTA 2.0 offers a structured vocabulary that can surface these diverging rationales and facilitate dialogue across professional boundaries.

3.3.2 Economy: Cost-Benefit Uncertainty and Coordination Gaps

Across all groups, economic concerns are central, but interpreted through different logics and institutional priorities. Several informants question whether digital health technologies deliver actual savings or merely displace costs. A common concern is that while national policies emphasise innovation and efficiency, implementation costs are often borne by frontline professionals without additional resources. As one regulatory informant noted:

> *'It sounds good that patients don't have to come to the hospital, but if a nurse has to spend two hours every Friday going through vital parameters on a screen, maybe we haven't actually saved anything on labor.'* (I1, Professional and regulatory organization)

Innovation and digitalisation actors often frame economy in terms of long-term return on investment, scalability, and cost-effectiveness at a system level. However, several also highlight the lack of coordination between institutions, which results in inefficiencies and lost opportunities for collective procurement:

> *'Each department or region often purchases its own equipment, like full-body scanners. If we coordinated better, regionally or nationally, we could probably save money by buying in bulk.'* (I2, Innovation and digitalization)

Healthcare professionals take a more pragmatic stance. Their focus lies on hidden costs: time spent on documentation, managing new tasks/invisible work, and disruptions to clinical routines, costs that are rarely acknowledged in budget models. One healthcare professional (I16) expressed, that time spent navigating new systems is rarely compensated or offset by workload reduction.

Educators and academics draw attention to challenges of continuing education. They note that integrating digital health into already packed curricula requires trade-offs: *'There are already so many mandatory themes, it's not easy to create space for new things, even when they're important'* (I5, Academics and educators)

Finally, regulatory and policy-oriented informants highlight the lack of frameworks to evaluate economic impact across institutional boundaries. Several stress the need for cross-sector models that consider not only direct financial savings, but also implications for staffing, service quality, and equity.

Taken together, these perspectives suggest that while 'economic value' is widely invoked, its definition is contested. For some, it implies future efficiency; for others, it highlights immediate strain. HTA 2.0 offers an opportunity to make these tensions visible by encouraging assessment practices that include both local workload and system-level return, fostering more realistic and accountable decision-making.

3.3.3 Environment: Sustainability and Technology Lifecycles

Despite increasing political attention to green transitions, environmental sustainability remains a notably marginal theme in most formal assessments of digital health technologies. Across the interviews, informants generally agree that environmental impacts are rarely prioritized in procurement, implementation, or professional training.

Innovation and digitalisation actors describe a lack of lifecycle thinking, where technologies are introduced without consideration of durability, upgradeability, or waste. One consultant expressed frustration with premature obsolescence: *'We replace entire systems after just a few years, that can't be sustainable'* (I9, Innovation and digitalisation)

Regulatory and professional actors echo this concern, pointing to the need for more structured integration of environmental criteria. One informant highlighted international models as more advanced in this regard:

> *'The Canadian HTA model is more flexible as sustainability and ethics are included. I think that could have real impact if applied at a hospital level… But if [sustainability and ethics] became more of a general mindset, that's where the potential lies and where it could make a real impact.*' (I1, Professional and regulatory organization)

Educators and academics suggest that sustainability could be embedded in training, not just as a technical theme but as a component of ethical and professional awareness. They propose linking environmental considerations to broader discussions about responsible innovation and resource use. By contrast, healthcare professionals rarely mention environmental issues unprompted, reflecting the acute time pressures and prioritization of patient care. As such, sustainability often becomes an invisible dimension in day-to-day healthcare practices. Taken together, the interviews suggest that while sustainability is recognized as important, it is rarely operationalized. HTA 2.0 could help surface environmental concerns by treating them as a legitimate dimension of assessment**,** particularly when linked to cost, durability, and responsible use of public resources. In this way, the framework may help move sustainability from rhetorical commitment to practical consideration.

3.3.4 Patient and Citizen: Digital Divide and Relational Concerns

Interviewees note that patients' digital competencies vary widely, and that healthcare professionals are increasingly expected to support, guide, and assess patients in their use of digital services. Several respondents raise concerns that digital tools, while designed to optimize processes, risk undermining relational aspects of care if not implemented thoughtfully. One regulatory informant expressed frustration that the patient perspective is often instrumentalized:

> *'The patient perspective is often poorly addressed by health tech companies. If it is considered, it's usually for marketing purposes. But we're more interested in whether the technology truly benefits the patient or helps the healthcare system save resources. Often, someone has come up with a clever idea they want to profit from, and the patient view gets lost in the process.'* (I1, Professional and regulatory organization)

Care work is repeatedly described as relational and ethically grounded. One innovation consultant emphasizes that new technologies inevitably reshape this dynamic:

> *'When you work in healthcare, you carry a deep relational responsibility toward the patient in front of you. Whether you are making clinical decisions or supporting basic care needs, the interaction is grounded in respect and integrity. Introducing new technologies into this space is never neutral. If professionals are equipped to reflect on how digital tools shape these encounters, the quality of care can be preserved.'* (I11, Innovation and digitalization)

These reflections point to a need for more systematic attention to the patient experience, not merely in terms of usability but as part of the ethical and relational fabric of care. Educators emphasize the importance of preparing students to adapt technology use to individual patients and maintain empathy in digital encounters. One educator (I5) noted that digital competence includes understanding how patients engage differently with tools and how those shapes clinical relationships.

Innovation actors often refer to patient feedback in terms of usability studies, interface design, or quantitative evaluations. One consultant, however, noted that such insights rarely address deeper experiences of care: '*Users may say the system works well, but that doesn't tell us how it affects trust or conversation in a patient consultation setting.'* (I3, Innovation and digitalization)

Regulatory informants call for more structured involvement of patients in assessment and policy development, warning against technologies that unintentionally widen the digital divide. As digitalization increases, the inclusion of diverse patient perspectives is seen as essential to ensuring equity and responsiveness. Taken together, the data show that while the value of the patient perspective is widely acknowledged, its interpretation varies across groups. HTA 2.0 can serve to make these differences explicit, supporting dialogue about how technology affects not just outcomes but also care relationships, trust, and inclusion.

3.3.5 Organization: Structural Constraints and the Role of Leadership

Organizational conditions strongly shape whether technologies succeed or fail in practice. Across interviews, informants consistently highlight that digitalization efforts are undermined when time, training, and communication are insufficient. Importantly, the success of implementation is not only dependent on the technology itself, but on leadership engagement, staff involvement, and the ability to articulate the rationality behind change. Healthcare professionals describe fragmented leadership and a lack of clear communication. One nurse emphasized the consequences of top-down rollouts: *'You don't implement by dropping a new system in our inbox. We need to know why and how'* (I12, Healthcare professional).

Many clinicians note that organizational support is often inconsistent, particularly when new technologies are introduced without sufficient planning, time, or follow-up. This results in resistance and frustration.

Innovation and digitalization actors view organisations as key levers for transformation. They speak of change management strategies, leadership metrics, and implementation roadmaps. Yet several acknowledge that this perspective is often misaligned with clinical realities. One consultant notes:

> *'If we don't explain why we are implementing this technology, people just see it as an annoying system disrupting their everyday work.'* (I6, Innovation and digitalization)

Another pointed out the absence of clearly defined roles for sustaining implementation (I9, Innovation and digitalisation). Educators and academics highlight that organizational support is crucial for enabling digital competence development. They argue that time for learning must be built into the system and that digital upskilling should not rely on individual initiative alone. Instead, structural enablers and recognition are necessary to ensure that digitalization becomes a supported part of professional development. Representatives of regulatory and professional organizations focus on governance, coherence, and accountability. Several informants express concern that without formal structures to assign responsibility for digital change, implementation efforts become fragmented or unsustainable. They stress the importance of aligning initiatives across levels to avoid duplication and inefficiencies. Taken together, these perspectives reveal that digital transformation depends not only on tools and strategies, but also on organisational readiness and distributed responsibility. HTA 2.0 may help clarify these dynamics by making visible the conditions that shape implementation, not just what technologies do, but what it takes to make them work.

3.3.6 Ethics: Dilemmas and Decision Blind Spots

Ethical aspects are frequently described as present but insufficiently addressed in formal assessments. Across the interviews, informants agree that ethical dilemmas, ranging from surveillance and data ownership to opaque decision-making in AI, are highly relevant in everyday practice but often remain underexamined. Healthcare professionals tend to frame ethics as something embedded in daily practice, often under time pressure and operational stress. One nurse captures this tension succinctly: *'Sometimes I'm not sure if I'm doing the right thing, registering or caring.'* (I13, Healthcare professional). This reflection illustrates how ethical judgment is exercised not only in grand decisions but in small, routine choices that balance professional duty and human presence. Innovation and digitalisation actors are increasingly attentive to ethical issues like algorithmic bias, transparency, and unintended consequences. However, ethics is often addressed too late in the process: *'We talk about ethics when the system is live, but maybe we should do it earlier.'* (I7, Innovation and digitalisation)

Another informant mentions that it is important to inform and prepare healthcare professionals when AI systems are implemented, otherwise it can affect their trust to a new system:

> *'You need to prepare the staff if a technology like AI is coming to their department. You can't just say: here's an artefact, a closed black box, and no one knows what it does.'* (I2, Innovation and digitalisation)

Educators and academics emphasise ethics as a transversal competence and an integral part of professional identity. They advocate for embedding ethical reflections into all stages of training, not as an isolated topic but as part of critical thinking and decision-making in practice. Professional and regulatory actors stress the need for clearer frameworks and procedures to evaluate ethical implications during procurement and approval. They express concern that ethical questions are often overlooked due to the absence of formal accountability mechanisms or relevant institutional routines. Taken together, these insights reveal that while ethical concerns are deeply felt across roles, they are not yet structurally integrated into assessment or implementation practices. HTA 2.0 can help address this gap by treating ethics not as an external constraint but as a legitimate and necessary dimension of technology assessment, linked to everyday dilemmas, institutional responsibilities, and anticipatory governance.

3.3.7 Diverging Perspectives Across Professional Groups: Can HTA 2.0 Support Cross-Professional Dialogue?

The preceding sections have shown how the six dimensions of HTA 2.0 resonate differently across professional groups. While the same themes recur, their interpretation, and prioritisation vary depending on institutional context, practical tasks, and professional roles. For some, technology represents systemic efficiency; for others, it introduces moral tensions, hidden costs, or relational disruptions.

These divergences do not reflect misunderstanding or resistance, but rather the multiple logics through which digital technologies are assessed in real-world settings. *Table 1* summarises how each group foregrounds different rationales and identifies where alignment or friction tends to occur. Several informants also reflect across domains, pointing to hybrid roles and emerging cross-professional awareness.

Group	Primary Rationales	Areas of Alignment or Tension
Innovation and Digitalisation	Focus on implementation capacity, scalability, and system-level efficiency	Possible tension arises when solutions lack clinical anchoring
Healthcare Professionals	Emphasis on usability, time pressure, and quality of care	Risk of resistance if implementation is top-down or adds workload
Academics and Educators	Promotes critical reflection, competence development, and pedagogy	Potential controversies between critical reflection and efficiency
Professional and Regulatory Organisations	Attention to evidence, equity, and cross-sector governance	Emphasize coordination and standardization while possibly overlooking practical situatedness

Table 1: Diverging Rationales and Tensions Across Health Actor Groups

What emerges is not a need for consensus, but for structured ways to articulate and negotiate these perspectives. HTA 2.0 does not erase institutional difference, it gives it form. By surfacing tensions that are often tacit, the framework can serve as a common language for critical dialogue, enabling healthcare actors to reflect on what technologies do, not only in terms of function, but in how they shape practice, responsibilities, and care. In the following section, we explore how these findings inform the design of continuing education initiatives and what it would take to embed HTA 2.0 as a boundary object that supports collective sense-making in complex healthcare environments.

4 Discussions and conclusions

4.1 Synthesizing Empirical Findings

The findings indicate a broader shift in how healthcare actors perceive the role of continuing education. Rather than focusing merely on the use of technologies, participants describe education as a space for developing the capacity to enact technologies responsibly in context to interpret, adapt, and negotiate digital systems within complex organizational and ethical environments. The informants thus see continuing education not as an add-on to implementation, but as a mechanism for translating technological ambitions into workable and meaningful practices.

The empirical material highlights how practical constraints, such as limited time, unclear responsibilities, and fragmented processes, challenge the implementation of digital technologies in everyday healthcare. At the same time, the analysis reveals that classical

HTA domains fail to capture critical aspects that matter to professionals, particularly ethical dilemmas, and environmental concerns. These insights underscore that effective education must move beyond technical training to also engage with the institutional logics and interpretive differences that shape technology use. *Table 2* summarises how HTA 2.0 responds to these challenges by providing a framework that can both support structured assessment and enable shared reflection across professional boundaries.

Phase	Findings	Implications for HTA Education
The inductive coding of interviews	1. Impact of technology in practice 2. Structural barriers in the healthcare system 3. Digital competencies and critical tech literacy 4. Leadership and organizational change 5. Technology integration in everyday work	HTA education must address real-world constraints such as lack of time, resources, and training support. It can focus on critical reflection, hands-on assessment skills, and integration of clinical realities into assessment.
The deductive coding of interviews	Six core dimensions of HTA 2.0: Technology, Economy, Organization, Patient/Citizen, Ethics, Environment	HTA education should go beyond technical assessments and integrate all HTA 2.0 dimensions. Tools to evaluate technology across disciplines and multiple societal dimensions are needed.
Synthesis of the two analytical approaches	Our synthesis identifies educational challenges: contextualizing assessment models, navigating value tensions, and enabling reflective dialogue across professional roles.	HTA 2.0 can serve as both an assessment tool to improve implementation of digital health technologies and a learning tool that can scaffold continuing education.

Table 2: Synthesis of findings from inductive and deductive coding and resulting implications for HTA education.

4.2 HTA 2.0 as an Educational and Reflective Framework

The HTA 2.0 model offers a structured yet flexible framework for interdisciplinary assessment of digital health technologies. It is both a pedagogical scaffold and a reflective tool that can be applied in clinical, municipal, and educational contexts. Based

on our findings, we suggest that HTA 2.0 can function as a model robust enough to structure shared dialogue, yet adaptable to local priorities and professional roles. The next step is then to see in practice how that works out.

Interviewees express a shared need for more reflective and situated assessment models, but they also articulate diverging expectations shaped by individual experiences, different roles and professional identities. This suggests that educational interventions should not seek consensus, but instead foster dialogical spaces where tensions between managerial, clinical, technical, and pedagogical logics can be surfaced and negotiated.

4.3 Continuing Education as a Boundary Object

A central insight from our study is that continuing education can act as 'boundary object' (Star & Griesemer, 1989): it does not merely transmit information or skills but enables reflection, anticipation, and sense-making in complex and dynamic work settings. This aligns with frameworks such as Responsible Research and Innovation (Stilgoe et al., 2013) and critical proximity (Amanatidis & Børsen, 2024), which emphasize inclusion and responsiveness in technology governance. By embedding HTA 2.0 in professional education, practitioners gain tools to assess not only efficacy, but also the societal, ethical, and organizational implications of digital transformation.

4.4 Policy Context and Anchoring of HTA 2.0

Several policy frameworks emphasize the need for context-sensitive digital health implementation. At EU level, *Regulation 2021/2282* calls for harmonized HTA procedures that support evidence-based implementation adapted to national contexts (European Parliament and Council, 2021). The *WHO Digital Health Action Plan* stresses governance, digital literacy, and equity (WHO, 2022). Nationally, the *Danish Strategy for Digital Health* (Sundheds- og Ældreministeriet, KL & Danske Regioner, (2018), the *Resilience Commission*'s recommendations (Robusthedskommissionen, 2023), and the *Life Science Strategy* (Uddannelses- og Forskningsministeriet, 2023) highlight innovation, workforce optimization, and digital competencies as key priorities.

Despite these ambitions, it remains unclear how such frameworks are to be translated into concrete institutional practices. While values like sustainability, inclusion, and ethical responsibility are prominently featured in strategic language, their operationalization in professional education, technology assessment, or implementation guidance is often vague or lacking.

In this context, HTA 2.0 may offer one possible contribution. As a reflective and practice-oriented framework, it could support the operationalization of policy ambitions**,** but only if it is adapted to local conditions and embedded for example in continuing education for health actors. Such integration could support a more grounded and critical approach to

digital transformation in healthcare**,** but it would require political initiatives, institutional support, and ongoing dialogue across groups.

4.5 Methodological Limitations and Future Directions

A limitation of this study is the underrepresentation of frontline health professionals. This may have constrained our ability to capture how assessment frameworks resonate with day-to-day care practices. Future workshops should aim to include more voices of healthcare professionals, to ensure that education and assessment tools align with practical concerns. Moreover, the overrepresentation of informants with innovation roles may have skewed some findings toward strategic or optimistic framings. While their insights are valuable, broader inclusion could reveal further tensions and implementation barriers.

In conclusion, HTA 2.0 has the potential to support more inclusive, critical, and context-sensitive approaches to digital health assessment and education. Rather than serving as a one-size-fits-all model, it offers a flexible framework that can scaffold interdisciplinary dialogue and reflective practice. Its integration into continuing education, could help bridge the persistent gap between technological ambition, implementation reality, and ensure that digital transformation in healthcare remains responsive to both professional expertise and societal values.

Our findings indicate that healthcare actors tend to view continuing education in HTA as a strategic mechanism for bridging the gap between digital ambitions and clinical enactment, by fostering reflections, negotiations, and a shared language across professional boundaries.

Acknowledgements

This research was funded by the Danish *Ministry of Higher Education and Science* under the national Life Science Strategy programme '*Enhanced competence development and continuing education in life science'* (Styrket kompetenceudvikling og efteruddannelse inden for life science). The authors would like to thank all participating healthcare actors and institutional partners for their valuable contributions.

References

Amanatidis, A. and Børsen, T. (2024). Critical proximity in translating RRI. Journal of Responsible Innovation, 11(1), 2373508. https://doi.org/10.1080/23299460.2024.2373508

Boenink, M. and Kudina, O. (2020). Values in responsible research and innovation: From entities to practices. Journal of Responsible Innovation, 7(3), 450–470. https://doi.org/10.1080/23299460.2020.1787758

Børsen, T. and Mehlich, J. (2024). Responsible research and innovation and tertiary education in chemistry and chemical engineering. Digital Chemical Engineering, 12, 100169. https://doi.org/10.1016/j.dche.2024.100169

Børsen, T. (2025). Technology assessment and postnormal science. Futures, 166, 103515. https://doi.org/10.1016/j.futures.2024.103515

Carlile, P. R. (2002). A pragmatic view of knowledge and boundaries: Boundary objects in new product development. Organization Science, 13(4), 442–455. https://doi.org/10.1287/orsc.13.4.442.2953

Danske Regioner. (2022). Plan for at bekæmpe ventelister og personalemangel i sundhedsvæsenet. Available at: https://www.regioner.dk/services/nyheder/2022/september/danske-regioner-lancerer-plan-for-at-bekaempe-ventelister-og-personalemangel-i-sundhedsvaesenet

Doezema, T., Ludwig, D., Macnaghten, P., Shelley-Egan, C. and Forsberg, E.-M. (2019). Translation, transduction and transformation: Expanding practices of responsibility across borders. Journal of Responsible Innovation, 6(3), 323–331. https://doi.org/10.1080/23299460.2019.1676686

European Parliament and Council. (2021). Regulation (EU) 2021/2282 on health technology assessment and amending Directive 2011/24/EU. Official Journal of the European Union, L 458, 1–27. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32021R2282

Gaedt, L., & Pedersen, K. F. (2017). Værdien af sensorgulve på plejecentre: Slutevaluering. Teknologisk Institut, Center for Velfærds- og Interaktionsteknologi. Available at: file:///Users/miebasballejensen/Downloads/V%C3%A6rdien%20af%20sensorgulve%20p%C3%A5%20plejecentre%20(12).pdf

Hansen, J. Z., Dalgaard, T. N. and Andersen, M. (2022). Langsigtet økonomisk fremskrivning 2021: Vurdering af den finanspolitiske holdbarhed. DREAM. Available at: https://www.dreamgruppen.dk

Højgaard, B. and Kjellberg, J. (2017). Fem megatrends der udfordrer fremtidens sundhedsvæsen. København: KORA. Available at: https://www.regioner.dk/media/4812/fem-megatrends-der-udfordrer-fremtidens-sundhedsvaesen-kora-2017.pdf

Indenrigs- og Sundhedsministeriet. (2023). Robusthedskommissionens anbefalinger. Available at: https://ism.dk/Media/638336462586551242/Robusthed-Samlet-Rapport-TILG.pdf

Jensen, M. B. and Børsen, T. (2025). Input fra sundhedsaktører om 'efteruddannelse i teknologivurdering for sundhedsprofessionelle' (SUNDTEK): Delrapport. Aalborg University Open Publishing. Available at: https://vbn.aau.dk/da/publications/input-fra-sundhedsakt%C3%B8rer-om-efteruddannelse-i-teknologivurdering

Kidholm, K., Jensen, L. K., Johansson, M. and Montori, V. M. (2024). Telemedicine and the assessment of clinician time: A scoping review. International Journal of Technology Assessment in Health Care, 40(1), e3. https://doi.org/10.1017/S0266462323002830

KL. (2022). Velfærdsteknologi i Norden i en tid med mangel på arbejdskraft. Kommunernes Landsforening. Available at: https://www.kl.dk/media/zptoyo4j/velfaerdsteknologi-i-norden-i-en-tid-med-mangel-paa-arbejdskraft.pdf

Langstrup, H. and Gjødsbøl, I. M. (2023). Forskere: Teknologiske quickfix kan ikke alene lukke dødens gab. Altinget. Available at: https://www.altinget.dk/forskning/artikel/forskere-teknologiske-quickfix-kan-ikke-alene-lukke-doedens-gab

Levina, N. and Vaast, E. (2005). The emergence of boundary spanning competence in practice: Implications for implementation and use of information systems. MIS Quarterly, 29(2), 335–363. https://doi.org/10.2307/25148682

Oskarsen, J. S. and Bratteteig, T. (2024). 'We kind of have to do our job alongside the digitalization' – on working with continuously changing tools. In: Proceedings of the 22nd European Conference on Computer-Supported Cooperative Work. https://doi.org/10.48340/ecscw2024_ep06

Owen, R., von Schomberg, R. and Macnaghten, P. (2021). An unfinished journey? Reflections on a decade of responsible research and innovation. Journal of Responsible Innovation, 8(2), 217–233. https://doi.org/10.1080/23299460.2021.1948789

Rip, A., Kulve, H.t. (2008). Constructive Technology Assessment and Socio-Technical Scenarios. In: Fisher, E., Selin, C., Wetmore, J.M. (eds) Presenting Futures. The Yearbook of Nanotechnology in Society, vol 1. Springer, Dordrecht. https://doi.org/10.1007/978-1-4020-8416-4_4

Riiskjær, E. (2014). Patienten som partner – en nødvendig idé med ringe plads. I S. Tjørnhøj-Thomsen & H. Kristensen (red.), Patientinddragelse i sundhedsvæsenet. København: Munksgaard.

Robinson, D., Winickoff, D. and Kreiling, L. (2023). Technology assessment for emerging technology: Meeting new demands for strategic intelligence. OECD Science, Technology and Industry Policy Papers, No. 146. https://doi.org/10.1787/e738fcdf-en

Shemesh, B., Coughlan, E. and Horton, T. (2025). Tech to save time: How the NHS can realise the benefits. The Health Foundation. Available at: https://www.health.org.uk/publications/long-reads/tech-to-save-time

Star, S. L. and Griesemer, J. R. (1989). Institutional ecology, 'translations' and boundary objects: Amateurs and professionals in Berkeley's museum of vertebrate zoology, 1907–39. Social Studies of Science, 19(3), 387–420. https://doi.org/10.1177/030631289019003001

Stilgoe, J., Owen, R. and Macnaghten, P. (2013). Developing a framework for responsible innovation. Research Policy, 42(9), 1568–1580. https://doi.org/10.1016/j.respol.2013.05.008

Sundheds- og Ældreministeriet, KL & Danske Regioner. (2018). National strategi for digital sundhed 2018–2022. Available at: https://sundhedsdatastyrelsen.dk/Media/638641635032899134/Strategi%20for%20digital%20sundhed%202018_2022.pdf

Sundhedsmonitor. (2024). Sundhedspersonale flygter: Opsigelser er steget med 50 procent på fire år. Available at: https://sundhedsmonitor.dk/debat/art9863773/Sundhedspersonale-flygter.-Opsigelser-er-steget-med-50-procent-på-fire-år

Uddannelses- og Forskningsministeriet. (2023). Styrket kompetenceudvikling og efteruddannelse inden for life science. København: Uddannelses- og Forskningsministeriet. Available at: https://ufm.dk/publikationer/2023/styrket-kompetenceudvikling-og-efteruddannelse-inden-for-life-science

Ugeskriftet. (2018a). Forskningschef advarer: Tænk jer om, når I tager ny sundhedsteknologi i brug. Ugeskrift for Læger. Available at: https://ugeskriftet.dk/nyhed/forskningschef-advarer-taenk-jer-om-nar-i-tager-ny-sundhedsteknologi-i-brug

Ugeskriftet. (2018b). Gode råd til digitalisering af det danske sundhedsvæsen. Ugeskrift for Læger. Available at: https://ugeskriftet.dk/debat/gode-rad-til-digitalisering-af-det-danske-sundhedsvaesen

World Health Organization. (2022). Regional digital health action plan for the WHO European Region 2023–2030. WHO Regional Office for Europe. Available at: https://www.who.int/europe/publications/i/item/EUR-RC72-5

Gaia women* garden: Co-Creating a space for transformative learning on bio-/diversity

Sandra Karner, David Steinwender, Anita Thaler

Interdisciplinary Research Centre for Technology, Work and Culture, Austria

DOI 10.3217/978-3-99161-062-5-013, CC BY 4.0

Abstract. This paper explores how biodiversity-related learning emerged in the GAIA Gartenberg women's garden in Graz, Austria. Drawing on Science and Technology Studies (STS), we analyse the garden as a socio-material learning environment in which knowledge, agency, and ecological relations are co-produced. Using reflection protocols, field notes, interviews, and Systematisation of Experience workshops, we explore how participation unfolded throughout the gardening season in 2024. The findings show that participation became possible for women in precarious life situations because enabling infrastructures such as land access, childcare, and translation were in place. Relational practices within the group fostered a cohesive and supportive community. Situated learning emerged through embodied, biographically rooted and culturally grounded engagements with plants, soil, insects, and food. More-than-human care practices reshaped participants' ecological sensibilities, and over time, processes of self-organisation, empowerment, and civic agency developed. The study states that transformative learning arises from socio-material and multispecies relations, and community gardens may be considered situated infrastructures of care and co-production that enable inclusive transformative learning.

Introduction

In recent years, growing attention has been put on educational initiatives that aim at addressing ecological crises, including climate change and biodiversity loss. These initiatives often seek to promote behavioural change by encouraging more sustainable lifestyles and consumption patterns. However, despite increased awareness and concern, a persistent discrepancy remains between individuals' values and their actual behaviours. This phenomenon is referred to as the knowing–doing gap, or the attitude–behaviour/value–action gap, which has been extensively explored in environmental psychology (Festinger, 1957; Bentler et al., 2023). STS and feminist scholars have long shown that such knowing–doing gaps cannot be understood solely as cognitive failures but must be analysed through the socio-material and affective conditions that shape

possibilities for action (Haraway, 1988; Suchman, 2007; Latour, 2004). These perspectives highlight how knowledge, agency, and behaviour emerge through relations among bodies, infrastructures, tools, institutional arrangements, and more-than-human actors. From an STS perspective, the focus shifts from individual deficits to the relational, material, and political conditions that enable or obstruct engagement with ecological issues. This includes structural inequalities, institutional path dependencies, and symbolic orders that shape who is able to participate in sustainability initiatives, on what terms, and with which forms of knowledge matter (Jasanoff, 2004). It also includes multispecies relations, e.g. with plants, insects, soil, that shape learning environments and practices (Puig de la Bellacasa, 2017; van Dooren et al., 2016; Houston et al., 2018).

Against this backdrop, this article investigates the Bio-/Diverse Edible City Graz case study within the Horizon Europe research project PLANET4B[1]. The study employed a participatory action research methodology and initiated learning communities at two interconnected scales: At the meso-level, a policy learning community was formed involving stakeholders from municipal administration, education, environmental sectors, social work, and the arts. At the micro-level, which represents the focus of this article, a community garden ('GAIA Gartenberg') was co-created by and for women* from diverse backgrounds, many of whom experience intersecting forms of marginalisation. Our analysis focuses on how biodiversity-related learning becomes possible when rooted in everyday, embodied practices; how socio-material environments co-produce agency, belonging, and ecological attentiveness; and how community gardening can generate forms of response-ability (Haraway, 2016) that expand participants' sense of what they can know and do.

Inequalities in Access to (Edible) Urban Green Space

Unequal access to urban green spaces reflects not just differences in physical availability but also structural inequities embedded in planning processes, socio-economic constraints, and symbolic orders of belonging (Anguelovski, 2013; Rigolon, 2016). Such inequalities are co-produced by infrastructures, governance arrangements, and cultural norms that privilege particular forms of participation and ecological knowledge (Jasanoff, 2004; Wynne, 1996).

Women, migrants, and residents with limited financial or linguistic resources often face barriers not only to accessing green spaces but also to feeling authorised to shape them (Kaijser & Kronsell, 2014). These dynamics are crucial for biodiversity learning: they influence whose experiences are recognised, which practices count as legitimate, and whose environmental relations inform urban transitions.

[1] https://planet4b.eu PLANET4B received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101082212

The Bio-/Diverse Edible City: An anchor for just transitions?

While visions of Edible Cities often emphasise self-sufficiency, greening, and health, they may obscure the social, political, and material dynamics that shape who benefits from these interventions (Säumel et al., 2019). From an STS perspective, such visions operate as socio-material imaginaries that mobilise particular urban futures and delineate who is expected to carry the responsibilities for care, maintenance, and ecological stewardship.

We expand the Edible City framework by foregrounding the 'bio' (biodiversity) and 'diverse' (inclusion, plurality) dimensions. This aligns with feminist and decolonial STS perspectives in which ecological practices are understood as situated, relational, and shaped by everyday forms of care (Puig de la Bellacasa, 2017). In this framing, the edible city becomes not a technical intervention but a socio-material and multispecies learning assemblage in which justice, knowledge, and care are negotiated.

Transformative Learning

Transformative learning (TL) has emerged as a critical pedagogical approach (Mezirow, 1990; 2000), yet its predominantly cognitive orientation may be enriched by STS perspectives that foreground how transformations are produced through situated, embodied, and materially mediated encounters. This is what Haraway (1988) calls situated knowledges, which emerge from specific positions, practices, and relations rather than abstract cognition alone. Her concept of situated knowledge shows how learning is grounded in partial, embodied experience—reflected in participants' sensory, cultural, and emotional engagements with plants, insects, and soil.

Jasanoff's (2004) co-production framework illuminates how learning processes intertwine with infrastructures, governance, and material arrangements.

Puig de la Bellacasa's (2017) work on care adds further analytic depth, framing learning as the unfolding of ethical and material entanglements with both human and more-than-human actors. Care here is not an add-on but a condition that enables knowledge, relations, and ecological attachments to grow (see also Houston et al., 2018). These perspectives reposition TL as a relational, socio-material process in which cognitive shifts are inseparable from changes in practice, sensibility, and response-ability (Haraway, 2016)

Community gardens as socio-material learning environments

Community gardens have increasingly been recognised as fertile ground for transformative learning that engages individuals not only cognitively, but also emotionally, socially, and ethically. These spaces foster forms of experiential learning that connect people to local ecologies, food systems, and community relationships, often catalysing shifts in perception, values, and behaviour (Mezirow, 2000). Community

gardens have been theorised as socio-material learning spaces where knowledge emerges through embodied activity, multispecies encounters, and collective experimentation (Pudup, 2008; Guitart et al., 2012). From an STS perspective, community gardens can be understood as socio-material *assemblages* (Deleuze & Guattari, 1987 in Müller, 2015) in which humans, tools, plants, insects, soils, and infrastructural arrangements co-constitute the conditions of possibility for action, collaboration, and learning. In these assemblages, knowledge and agency emerge not from individual actors alone but from the dynamic interactions among material arrangements, more-than-human organisms, and social relations.

This aligns with scholarship showing how civic and grassroots initiatives enact alternative environmental futures (Ghose & Pettygrove, 2014). In such settings, hands-on practices, shared labour, and collective decision-making enable participants to develop critical awareness of social and environmental injustices while simultaneously building practical skills and ecological literacy (Aiken, 2016; Egerer et al., 2019). Community gardens thus act as transformative learning environments where knowledge is co-created through embodied interaction with the land and other community members (Sipos et al., 2008). Consistent with this, the GAIA Gartenberg garden became a place where biodiversity was learned through touch, care, sensory experience, and multispecies entanglements.

Positioning community gardens in this way supports our analysis of the GAIA Gartenberg case as a socio-material and multispecies learning environment in which participants developed situated ecological knowledge, emotional attachments, and emerging civic agency.

Methodology

Our case study employed a participatory action research (PAR) design that combined collaborative garden co-creation with regular research interventions and ongoing qualitative evaluation. The methodological approach was informed by TL theory and complemented by STS concepts of co-production and care.

Case Study Description

The citizens' LC was implemented between March and September 2024 in Graz, Austria, as part of the PLANET4B project. The case study centred on the co-creation of the GAIA Gartenberg women's community garden, developed in close collaboration between IFZ researchers, gardeners from Forum Urbanes Gärtnern (FUG), and 10–15 participating women*. Meetings took place weekly on Fridays for 3 hours, regardless of the weather. Most sessions were held on-site, and some were held at a nearby community centre during periods of bad weather.

The garden space was intentionally designed as a women*-only environment, co-facilitated by pedagogically trained female gardeners from FUG, and guided by principles of brave spaces (Arao & Clemens, 2013), accessibility, and care. The weekly meetings combined gardening activities, collective decision-making, reflection exercises, shared meals, and structured research workshops on biodiversity, food systems, and diversity.

e conceptualised the GAIA Gartenberg as a socio-material setting in which knowledge, relations, and forms of agency were co-produced. Following Jasanoff (2004), we understand co-production as the intertwined making of social order, knowledge, and material arrangements. Rather than treating the garden as a neutral backdrop, we approached it as a dynamic assemblage of people, materials, practices, and norms that actively influenced the learning process.

To support learning processes and generate empirical material, a series of structured research interventions was embedded into the gardening process. These activities were designed to elicit experiential knowledge, stimulate reflection, and anchor biodiversity-related issues in everyday practice. The first activity was an '*experience-stroll*' through the future garden site. It encouraged participants to articulate prior gardening experiences, expectations, and cultural connections to plants and food. As the garden developed, further interventions included a community-mapping (Taliep & Ismail, 2023) exercise to identify needs, barriers, and opportunities for future use; socio-scientific inquiry workshops (e.g. on apple varieties; Zeidler & Kahn, 2014) to explore biodiversity, seasonality, labour conditions, and food system dynamics. Another workshop on diversity, accessibility, and inclusion was held to connect personal experiences with broader questions of justice.

Co-creative processes were a defining feature of the study design. Participants developed the garden plan together, drawing on their collective knowledge, preferences, and practical considerations. A thematic Milpa or 'Three Sisters' bed (inspired by Kimmerer, 2013) was jointly established, and the garden was collectively named 'GAIA Gartenberg'. These co-creative practices reflected an STS-informed commitment to epistemic pluralism, integrating practical, cultural, sensory, and situated expertise alongside scientific or technical knowledge (Norström et al., 2020; Klein, 2004; Lang et al., 2012).

All weekly meetings were structured along check-ins, gardening work, shared meals, reflective discussions, and check-outs. As such, they functioned simultaneously as pedagogical tools and as integral components of the research design, enabling the production of rich qualitative insights into learning, collaboration, and the socio-material dynamics shaping participants' engagement with biodiversity.

Data Sources

Our research relied on four qualitative data sources:

(a) Reflection protocols: After each Friday session, open-ended reflection protocols were completed by the FUG facilitation team. These protocols captured general observations, group dynamics, critical incidents, and emerging learning moments. They were not based on structured templates but followed a qualitative memo style that allowed for the flexible capturing of emerging issues.

(b) Researcher field notes: IFZ researchers maintained open, descriptive field notes, documenting interactions, decision-making processes, material challenges, and emotional or relational dynamics. Notes were recorded during or immediately after sessions.

(c) Individual follow-up interviews: Seven individual ex-post interviews were conducted with women* in early spring 2025, each lasting approximately 60 minutes. The interviews followed a semi-structured guide and focused on participants' reflections on learning processes, experiences of co-creation, encounters with biodiversity, perceptions of agency and belonging, and expectations for the future development of the garden. Participation was voluntary, and the sample size reflects availability and willingness rather than a strategic sampling.

(d) Systematisation of Experiences (SoE): At the end of the gardening season, four SoE workshops (Herout & Schmid, 2015) were conducted with the women*, the research team and members of the policy LC. This collective reflection exercise generated structured insights into how participants interpreted key moments, challenges, and turning points in the co-creation of the garden. The SoE workshops produced a collaboratively developed narrative of the group's learning journey, which served both as data and as a validation tool for preliminary interpretations.

Together, these four data sources provided complementary insights into how learning processes were co-produced through social, material, and affective engagements throughout the gardening season.

Data analysis

We analysed the four data sources outlined previously and followed an iterative, interpretive, and STS-informed approach. Rather than coding for predefined categories, we traced how meanings, practices, and relationships emerged across the season.

Analytically, we moved iteratively across these materials, guided by feminist and STS concepts of situated knowledge (Haraway, 1988), care (Puig de la Bellacasa, 2017)and *response-ability* (Haraway, 2017), and co-production (Jasanoff, 2004). We focused on: embodied and affective learning, the negotiation of interpersonal relations, socio-material

entanglements with plants, insects, tools, and infrastructures, and the emergence of care, agency, and collective responsibility.

Results

The empirical findings illustrate five interconnected dynamics: (1) how material, institutional, and social infrastructures enabled participation; (2) how a diverse group of women* gradually became a cohesive community through shared practices and mutual trust; (3) how experiential and culturally situated learning unfolded through workshops, everyday gardening, and creative methods; (4) how participants developed new forms of more-than-human attentiveness and care through encounters with plants, insects, soil organisms, and ecological processes; and (5) how self-organisation, confidence, and civic agency emerged as participants began to imagine and enact future responsibilities for the garden.

Enabling Conditions as Co-Produced Infrastructures

Before establishing the *citizens LC* of the *Bio-/Diverse Edible City Graz*, we began with a careful exploration of the setting – including the social context, spatial characteristics, and related factors and participants' needs.

To ensure low-threshold participation in the citizens LC, we needed to establish enabling conditions. Structural and emotional barriers, such as language, previous negative experiences with institutions, or unfamiliarity with environmental topics, were explicitly acknowledged and addressed through multilingual facilitation (German, English; and using Google Translate to Russian and Ukrainian), informal settings, and an appreciative approach to existing knowledge and practices.

Based on our reflections within the SoE we identified key resources at multiple levels:

(1) *Physical resources* included a dedicated garden plot (slightly remote, but easily accessible and big) provided by the city administration, access to essential infrastructure such as water, soil, compost, and tools, as well as storage space for materials, which were fundamental to ensuring that gardening and learning could take place consistently and safely.

(2) *Financial support* through the PLANET4B research project provided funding not only for material needs but also for process facilitation, childcare, translation support, and other activities that enhanced accessibility and continuity.

(3) *Personal and social resources* were essential for both understanding the local context and reaching out to potential participants. Local actors, such as community workers, neighbourhood organisations, and practitioners in community-based work, played a key role in identifying whom to approach, how, and where. They also served as trusted points

of contact within the community, owing to their established networks and trust-based relationships. These connections were crucial for engaging women* from diverse and often marginalised backgrounds and for lowering initial barriers to participation. As one participant expressed, *'I really felt received as if I belonged here […] I was welcomed warmly and openly.'* (EPI_W1_20022025)

(4) *Symbolic and institutional resources* included the absence of resistance from surrounding residents and the general goodwill of municipal actors, including those responsible for green space and urban development. An explicit political endorsement to foster urban social gardening, and the fact that the project encountered only little administrative or social obstruction, were themselves significant enabling conditions. In this sense, symbolic space - the room to experiment, fail, and grow without being overly scrutinised or instrumentalised - was just as important as physical space. All this allowed the project to maintain low-threshold participation while responding flexibly to participants' needs.

These enabling conditions allowed facilitators and participants to treat the garden as a site for learning rather than performance. In STS terms, these enabling conditions illustrate how material, institutional, and social arrangements co-produce the very possibility of inclusive participation.

Becoming a community: group formation, trust, identity

Considerable attention was given to creating relational safety, group-building and visioning. Thus, early engagement formats were intentionally designed as 'brave space' (Arao & Clemens, 2013), allowing participants to engage emotionally, share vulnerabilities, and navigate linguistic and cultural differences without fear. One woman captured this sense of grounding: *'Being welcomed, having a quiet moment at the start […] it helped me to slow down from everyday life.'* (EPI_4_25022025)

Routines and rituals, such as check-ins and check-outs, meals, and sharing personal stories and memories, helped to establish continuity and familiarity. As another participant recalled: *'We always had a check-in and a check-out […] it gave rhythm and helped us feel connected.'* (EPI_W5_11032025) These relational practices reflect Haraway's (1988) concept of situated knowledge: learning begins by locating oneself among others, within shared practices and affective ties.

A formative moment in community building occurred when the women* collectively decided to build the garden fence themselves. Initially, some suggested asking male relatives for help, yet the group chose to take the task into their own hands. The embodied experience of constructing the fence by driving posts into the earth, stretching wire, clearing stones etc. became a symbolic act of empowerment and claiming the space. Within the SoE this was highlighted a critical moment for several times, and its significance was emphasised in the interviews as well: *'Being able to hammer a fence*

post into the ground as a woman […] it was something I never imagined I could do.' (EPI_W4_25022025) Another woman remembered: *'We collected stones like ancient humans and stretched the fence […] and we women really managed it well.'* (EPI_W6_13032025)

This exemplifies infrastructuring as a co-productive process (Star & Bowker,2006): material arrangements (the fence) and social relations (confidence, trust, collective ownership) emerged together. As one participant summarised the developing community ethos: *'Through the working together, trust came naturally. You don't need the same language when your hands are doing the same task.'* (EPI_W6_13032025)

By the end of the season, a strong sense of collective identity had formed. Reflection notes from the second community mapping at the end of the harvesting season show how the garden became a place of belonging: *'When I think of the garden year, my body feels warm […] the place has become familiar.'* (CM2_follow-up reflections_25102024) This affective attachment resonates with Puig de la Bellacasa's (2017) understanding of care as an ongoing cultivation of relations, not only among humans but also toward place and the more-than-human, and with gratitude for the gifts the garden offered.

Situated learning practices: experience walks, workshops, mappings

As trust solidified, learning processes became increasingly situated, experiential, and co-created. Rather than introducing biodiversity and social diversity as abstract topics, facilitators grounded them in sensory experience, embodied practices, and participants' lived histories. This approach reflects Haraway's argument that knowledge emerges from partial perspectives located in experience.

Activities such as a '*nature experience stroll'* served as relational entry points into the topic of biodiversity, encouraging participants to share personal memories related to plants, food, and places from their own lives. These moments were not framed as didactic tools, but as openings for meaning-making rooted in lived experience. Reflection notes describe how 'the format enabled a form of learning through place: knowledge was not transmitted abstractly but anchored bodily, sensually and socially' (NES_reflection notes_22032024). Women* articulated connections between biodiversity and their own biographies: *'For me a garden means connecting and communicating with the earth and the plants. and seeing what grows from my hands'* (NES_reflection notes_22032024).

In the socio-scientific issues (SSI) workshop on apples, participants evaluated varieties based on taste, economic aspects, ecological issues, and cultural resonance. As reflection notes highlight: *'The diversity of apples surprised me; each taste led to a different discussion'* (SSI_WS1_reflection notes_26072024). Another participant summarised: *'Every workshop fed my mind; I learned so much from the others.'* (EPI_W3_Datum) This illustrates how ecological knowledge emerged through collective reasoning situated within everyday constraints and values.

Learning also unfolded through weekly routines. As one participant explained: *'We always did a check-in and a check-out […] during the break everyone brought something to eat; we talked about what we had cooked from last week's harvest and shared recipes in the WhatsApp group.'* (EPI_W5_11032024) Through such practices, knowledge was co-produced through communicating, doing, tasting, sharing, and reflecting together.

The diversity workshop further illustrated situated learning by making linguistic and cultural plurality visible: *'Hearing the word 'garden' spoken in twenty languages created a moment of pride and curiosity.'* (DWS_reflection notes_06092024) Here, difference became an epistemic resource rather than a barrier, aligning with feminist STS approaches that value plurality and relationality (Haraway 1988).

More-than-human and care practices

A central aspect of learning in the GAIA garden was the cultivation of more-than-human relations. Through touching, observing, and regular interaction, participants related to plants, insects, soil organisms, and seasonal rhythms. Haraway's notion of response-ability (2016) offers a useful conceptual lens here: learning involved becoming capable of responding to the needs and signals of non-human others.

Participants articulated this shift explicitly. One woman described learning new forms of attentiveness: *'Learning how much water each plant needs made me feel responsible for them […] like they depend on us.'* (EPI_W6_13032025) Another emphasised temporal ethics: *'The garden taught me patience. You cannot rush a plant. You have to care for it and wait.'* (EPI_W3_20032025)

Observations of insects also changed: *'When the flower meadow behind the garden emerged, it changed a lot visually and surely attracted many more insects.'* (EPI_W5_11032025) A particularly illustrative example of transformed relations to nature comes from a participant who spoke about overcoming her long-standing fear of insects. She explained that before joining the project, she could not sit on grass because she was afraid something might crawl onto her and would 'jump' whenever she encountered insects. Through repeated encounters in the garden and reassurance from others, this fear gradually diminished: *'This fear has been gone […] now I'm like, okay, it's okay.'* She also noted how observing other women calmly brushing insects away, and seeing children move freely and unafraid in the garden, helped her reframe these interactions: *'Looking at that, I thought it's not that big a deal to be in nature.'* (EPI_W7_11032025).

An even more complex example of ethical and emotional engagement with the more-than-human emerged in relation to the garden's recurring snail infestation. One participant described how she absolutely refused to kill snails, expressing strong discomfort and moral resistance: she carried hundreds of them by hand 'far up into the forest' rather than harming them. She explained that she could not kill a living being without certainty that it would suffer 'not even a second,' and that contradictory advice

from others only deepened her unease. She noted that '*everyone was somehow avoiding the topic, everyone was unsure'* (EPI_W6_13032025), illustrating how the group collectively navigated the ethical ambiguity of multispecies encounters. This narrative captures how care, uncertainty, and ethical negotiation shaped participants' relations with more-than-human life, revealing the garden as a site where moral and practical worlds are co-produced.

Such insights show how the garden functioned as a multispecies contact zone, a space where ecological processes became perceptible through embodied engagement. What was once experienced as a threat became normalised through collective practice and shared presence, which illustrates how response-ability emerges not through instruction but through situated, relational exposure.

Care practices extended beyond plant cultivation into cooking and food preparation. Participants exchanged recipes, experimented with unfamiliar vegetables, and developed confidence in using produce from the garden. As one woman shared: *'I always saw kohlrabi on the counter but never knew how to make it […] then people cooked it at home and brought it, and when I tasted it I thought: oh, this is nice. Now I know how to make it.'* (EPI_W7_22032025)

In Puig de la Bellacasa's terms (2017), these examples illustrate how care is simultaneously affective, material, and epistemic; it is a mode of engagement that binds people, plants, insects, soil, spaces, and shared meanings together.

Emergent agency, self-organisation and the future-making

By mid-summer, participants increasingly articulated long-term visions for the garden and demonstrated growing confidence in collectively managing it. This transition from facilitated engagement to autonomous self-organisation marks a key outcome of the co-productive learning process.

In the community mapping follow-up, women* expressed diverse yet converging future imaginaries. One noted: *'Now I can imagine the garden in a new year, who is there, what I will do, what grows […] it has become part of my life.'* (CM2_follow-up reflections_25102024) Another reflected: *'Now I know I can grow vegetables at home, Mangold, tomatoes […] it comes fast, and I feel confident.'* (EPI_W7_11032025)

Discussions about founding a nonprofit association, which is a requirement for long-term stewardship of community gardens in Austria, further revealed emerging civic agency. As one participant stated: *'I want to be a role model […] to do my part so the next generation can also harvest.'* (EPI_W5_11032025) Another summarised the group's growing autonomy at the end of the growing season: *'We no longer need someone to tell us what to do in the garden, we already know how to organise ourselves and decide together what needs to happen.'* (CM2_follow-up reflections_25102024)

Participants also emphasised the emotional significance of the women-only space: *'The women's space is special. Here I discovered abilities I never believed I had.'* (CM2_follow-up reflections_25102024) Another described the garden as a personal refuge: *'This is my happy place. A space to feel free and share experiences.'* (CM2_follow-up reflections_25102024) These statements illustrate how empowerment, identity, and belonging were co-produced through socio-material and relational practices.

Finally, women* articulated increased confidence in broader social settings: *'I am no longer afraid to join new groups; I know now that people will treat me kindly.'* (EPI_W6_13032025) This shift signals how the GAIA garden functioned as a site of civic learning, enabling women* to imagine themselves as active contributors to urban ecological futures.

Taken together, these findings show how agency, responsibility, and future-making were co-produced through the intertwined dynamics of everyday practice, shared decision-making, and care for the space.

Discussion

The case study shows how the citizen LC operated as a situated experiment in co-producing infrastructures, knowledges, and subjectivities, rather than as a neutral educational intervention. Bringing Jasanoff's (2004) notion of co-production together with the work of Haraway (1988, 2016) and Puig de la Bellacasa (2017) enables us to understand the GAIA Gartenberg case as an example of how social relations, material arrangements, and ways of knowing were created and reshaped together.

First, the results demonstrate that enabling conditions were not pre-given but actively *infrastructured* (Star & Bowker, 2006) through the joint work of the PLANET4B project team and the women themselves, with support from municipal actors. Access to land, water, tools, storage, childcare, translation, and process facilitation provided a socio-material basis for women* in precarious life situations to participate. Participants' descriptions of feeling 'received' and 'welcomed warmly and openly' indicate that infrastructuring was simultaneously material and affective. This resonates with co-production in Jasanoff's sense: institutional commitments to urban gardening, funding streams, and garden infrastructures did not simply support an already-existing learning process. These aspects co-defined who could become a participant, which forms of knowledge were legitimate, and what futures could be imagined.

Second, the findings specify how situated knowledge (Haraway, 1988) was generated through embodied and culturally inflected practices. Activities such as the nature experience walk, the socio-scientific issues workshops, and community mappings anchored biodiversity in everyday experiences of taste, memory, and place. When women linked gardens to childhood memories, migration histories, or family recipes, they

enacted what Haraway calls *partial perspectives*: knowledges that are local, accountable, and entangled with biography rather than abstract universal truths. The multilingual setting and the ongoing translation support from facilitators and peers show that knowledge was not simply passed on but carefully built together across different languages, cultures, and experiences. In this sense, the GAIA garden became a site where epistemic authority was redistributed and where gardening expertise, sensory impressions, and everyday food practices were treated as legitimate contributions alongside scientific or policy-oriented perspectives.

Third, the project shows how practices of care were central to learning, aligning with Puig de la Bellacasa's (2017) understanding of care as affective, material, and epistemic at once. Care appeared in everyday activities, like watering plants, collectively managing the 'snail problem,' experimenting with new vegetables, cooking and sharing food, or gently supporting women who were initially anxious about insects or social exposure. These practices did more than maintain the garden: they produced attachments, responsibilities, and forms of attentiveness. The woman who carefully carried snails 'far into the forest' rather than killing them, and who later reflected on her uncertainty about how to avoid causing suffering, exemplifies how ethical and ecological questions became folded into everyday routines. Likewise, the participant who overcame her long-standing fear of insects by watching others brush them off calmly, and by observing how children moved unafraid through the garden, illustrates how *response-ability* (Haraway, 2016) is cultivated through repeated multispecies encounters rather than through moral injunctions alone.

Fourth, the study contributes to debates on transformative learning by showing how transformation unfolded as a socio-material, more-than-human process rather than as a purely cognitive shift. Across the season, women reported increased confidence in gardening, food preparation (e.g. fermenting vegetables and cooking with previously unfamiliar varieties), and joining new social groups. These changes were tethered to concrete practices: building the fence, co-designing the garden layout, reflecting on the value of varieties, or co-founding the association. The transition from relying on facilitation to articulating that they would not need someone to tell them what to do in the garden signals a shift in agency that is inseparable from the shared work of maintaining beds, negotiating responsibilities, and imagining future uses of the site. Rather than a sudden *disorienting dilemma* (Mezirow, 2000), transformation in this context unfolded gradually, as participants' sense of what they could do shifted through their ongoing engagement with tools, soils, plants, institutional actors, and one another.

Fifth, the case foregrounds the gendered and intersectional dimensions of co-production. The women-only setting functioned as a protected, *brave space* in Arao and Clemens' (2013) sense, where participants could experiment with new roles and practices without fear of ridicule or surveillance. Many women described discovering abilities 'I never believed I had' and naming the garden as a 'happy place' and refuge. For women affected

by migration, low income, care burdens and language barriers, the combination of spatial seclusion, female facilitation, and low-threshold entry points (food, children welcome, no prior expertise required) was crucial. From an intersectional STS perspective (Cranshaw, 1989), this highlights that inclusive edible city initiatives must address not only physical access but also symbolic safety, gendered power relations, and the time–care regimes that structure who can participate and when.

Finally, the discussion must also address tensions and limits. The project depended heavily on external funding, committed facilitators, and an unusually supportive municipal context. Infrastructuring, in this sense, is both enabling and fragile: if funding streams or political priorities shift, the carefully co-produced conditions for participation may erode. Moreover, while the case aimed to redistribute agency, it also relied on unpaid volunteer labour and emotional work by facilitators and participants—raising questions echoing critiques of neoliberal *responsibilisation* in community-based sustainability initiatives (Mayer, 2012; Rosol, 2017). The more-than-human dimension, though present, remained somewhat bounded by the immediate concerns of food, pests, and plant care; broader biodiversity politics, species loss, or contested land-use regimes could only be touched upon. These limitations underline that GAIA Gartenberg should not be read as a fully realised alternative but as a situated, partial experiment that opens particular possibilities while leaving other structural dynamics intact.

Taken together, the Bio-/Diverse Edible City Graz case shows that when co-production, situated knowledge, and care are taken seriously, community gardens can become laboratories for reconfiguring socio-ecological relations. The garden did not solve systemic inequalities, but it made them negotiable in new ways, allowing women to inhabit roles as gardeners, association founders, neighbours, and carers of plants and insects that previously seemed closed or risky. In doing so, the case adds empirical texture to STS debates about how small-scale, everyday practices can participate in the making of more just and liveable urban futures.

Conclusions

This article has examined how transformative learning about biodiversity and urban nature emerged in the GAIA Gartenberg women’s garden as part of the Bio-/Diverse Edible City Graz case. By bringing feminist and multispecies STS concepts into dialogue with transformative learning theory, we argued that learning in this context was co-produced through specific socio-material arrangements and care practices rather than through information transfer alone. The study's key contributions lie in three interrelated insights:

First, the case demonstrates that enabling participation for women in precarious life situations requires more than ‘inviting’ them into pre-existing initiatives. It demands

deliberate infrastructuring: securing land and basic infrastructure; providing translation, childcare, and respectful facilitation; and cultivating institutional goodwill that protects the project from bureaucratic friction. These infrastructures are not mere background conditions but active components of co-production, shaping whose knowledge counts and whose futures are thinkable in the edible city.

Second, the findings demonstrate how biodiversity-related learning becomes meaningful when it is anchored in situated knowledges and more-than-human relations. Experience walks, apple-tasting workshops, community mapping, and everyday encounters with plants and insects allowed women to weave ecological concerns into their own biographies, emotions, and routines. Through these entangled practices, response-ability and care extended from the human community to soil organisms, insects, and cultivated plants, illustrating Puig de la Bellacasa's (2017) claim that caring is simultaneously about maintaining worlds and learning to know with others.

Third, the GAIA Gartenberg case illuminates how community gardens can function as small but significant sites of future-making. The emergence of self-organisation, the founding of an association, and the articulation of long-term visions for the area demonstrate how participants began to see themselves as legitimate actors in urban socio-ecological transformations. The garden became an anchor for broader developments—such as planned orchard meadows and a community park—while also feeding back into policy discussions through the policy LC. In this way, the case renders visible how modest, local initiatives can ripple into wider governance arenas when supported by attentive facilitation and receptive institutions.

At the same time, we emphasise that these outcomes are context-specific and non-transferable in a simple sense. Replication elsewhere would require not only similar funding and municipal support but also careful attention to local histories, power relations, and more-than-human ecologies. Rather than offering a template, the Bio-/Diverse Edible City Graz case provides a situated example that can inspire other actors to ask: what infrastructures of care, what forms of co-production, and whose situated knowledges would be needed to enable comparable processes here?

For researchers and practitioners working at the intersection of STS, urban governance, and sustainability education, the case suggests that designing transformative learning environments means designing socio-material worlds: assembling infrastructures, relations, and practices that allow marginalised groups to experiment with new ways of knowing and acting. Future work could deepen this perspective by tracing longitudinally how such initiatives endure or transform once project support ends, and by engaging more explicitly with conflicts and frictions over land use, labour, or species priorities that inevitably accompany attempts to reconfigure urban socio-ecologies.

In sum, the Bio-/Diverse Edible City Graz case underscores that transformative learning in community gardens is not a method to be applied but an emergent process. It emerges

where shared infrastructures, situated knowledges and practices of care come together, creating conditions that allow different urban futures to be imagined and tried out, even if only temporarily.

Methodological Limitations

While the analysis highlights more-than-human relations, our empirical access to multispecies interactions was inevitably partial and mediated through human accounts, observations, and research interventions. This means that the more-than-human perspective remained largely inferred rather than systematically documented, reflecting a common challenge in ethnographic and participatory STS research. The focus, by design, remained on making biodiversity relevant to the everyday lives, priorities, and cultural frames of the participating women*, which itself can be understood as a necessary first step toward inclusive ecological literacy.

Acknowledgements

This research was funded by the European Union’s Horizon Europe research and innovation programme under grant agreement No. 101082212. We thank the women involved in the GAIA Gartenberg community garden for their dedication and enthusiasm. Our sincere appreciation also goes to our colleagues at Forum Urbanes Gärtnern and Community Centre Eggenlend for the fruitful and inspiring collaboration, as well as to the experts from Inspire, NaturErlebnisPark, ESC Medien Kunst Labor, and the University College of Teacher Education Styria for their valuable contributions to the implementation of the case study. Finally, we gratefully acknowledge the support of the Municipality of Graz, particularly the Department of Green Spaces and Waterways.

AI use disclosure

Parts of the writing process were supported by using AI tools. Perplexity.ai and ChatGPT (v4) were employed for literature searches, while ChatGPT also assisted in developing structural ideas for the results chapter and refining language and phrasing, based on author-provided arguments and partial drafts. All final texts were carefully revised and proofread by the authors. The interpretations and final decisions are solely the responsibility of the authors.

References

Aiken, G. T. (2016). Community transitions to low carbon futures in the Transition Towns Network (TTN). Geoforum, 74, 1–10. https://doi.org/10.1016/j.geoforum.2016.05.009

Anguelovski, I. (2013). From environmental trauma to safe haven: Place attachment and place remaking in three marginalized neighborhoods of Barcelona, Boston, and Havana. City & Community, 12(3), 211–237.

Arao, B., & Clemens, K. (2013). From safe spaces to brave spaces: A new way to frame dialogue around diversity and social justice. In L. Landreman (Ed.), The art of effective facilitation: Reflections from social justice educators (pp. 135–150). Stylus Publishing.

Bentler, D., Kadi, G., & Maier, G. W. (2023). Increasing pro-environmental behavior in the home and work contexts through cognitive dissonance and autonomy. Frontiers in Psychology, 14, 1199363.

Crenshaw, K. (1989). Demarginalizing the intersection of race and sex: A Black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. University of Chicago Legal Forum, 1989(1), 139–167.

Deleuze, G., & Guattari, F. (1987). A Thousand Plateaus: Capitalism and Schizophrenia. (B. Massumi, Trans.). University of Minnesota Press.

Egerer, M. H., Fairbairn, M., & Winkler, R. (2019). Urban gardens as spaces of citizenship. Local Environment, 24(4), 304–317. https://doi.org/10.1080/13549839.2019.1590322Angeles, I. T. (2023). Urban Gardening: A Catalyst for Women's Empowerment, Community Engagement, and Environmental Awareness. Community Engagement, and Environmental Awareness.

Festinger, L. (1957). A theory of cognitive dissonance. Stanford University Press.

Ghose, R., & Pettygrove, M. (2014). Urban community gardens as spaces of citizenship. Antipode, 46(4), 1092–1112.

Guitart, D., Pickering, C., & Byrne, J. (2012). Past results and future directions in urban community garden research. Urban Forestry & Urban Greening, 11(4), 364–373. https://doi.org/10.1016/j.ufug.2012.06.007

Haraway, D. J. (1988). Situated knowledges: The science question in feminism and the privilege of partial perspective. Feminist Studies, 14(3), 575–599. https://doi.org/10.2307/3178066

Haraway, D. J. (2016). Staying with the Trouble: Making Kin in the Chthulucene. Duke University Press.

Herout, V., & Schmid, C. (2015). Systematisierung von Erfahrungen: Ein Leitfaden für die Praxis. Wien: Verband Österreichischer Entwicklungsorganisationen (AG Globale Verantwortung).

Houston, D., Hillier, J., MacCallum, D., Steele, W., & Byrne, J. (2018). Make kin, not cities! Multispecies entanglements and 'becoming-world' in planning theory. Planning Theory, 17(2), 190–212. https://doi.org/10.1177/1473095216688042

Jasanoff, S. (2004). States of Knowledge: The Co-Production of Science and Social Order. Routledge.

Kaijser, A., & Kronsell, A. (2014). Climate change through the lens of intersectionality. Environmental Politics, 23(3), 417–433. https://doi.org/10.1080/09644016.2013.835203

Kimmerer, R. W. (2013). Braiding Sweetgrass: Indigenous Wisdom, Scientific Knowledge and the Teachings of Plants. Milkweed Editions.

Klein, J. T. (2004). Prospects for transdisciplinarity. Futures, 36(4), 515–526. https://doi.org/10.1016/j.futures.2003.10.007

Lang, D. J., Wiek, A., Bergmann, M., Stauffacher, M., Martens, P., Moll, P., Swilling, M., & Thomas, C. J. (2012). Transdisciplinary research in sustainability science: Practice, principles, and challenges. Ecological Economics, 79, 1–10. https://doi.org/10.1016/j.ecolecon.2012.04.017

Latour, B. (2004). Why has critique run out of steam? From matters of fact to matters of concern. Critical Inquiry, 30(2), 225–248. https://doi.org/10.1086/421123

Mezirow, J. (1990). How critical reflection triggers transformative learning. In J. Mezirow & Associates (Eds.), Fostering critical reflection in adulthood (pp. 1–20). Jossey-Bass.

Mezirow, J. (2000). Learning as Transformation: Critical Perspectives on a Theory in Progress. The Jossey-Bass Higher and Adult Education Series. Jossey-Bass Publishers, 350 Sansome Way, San Francisco, CA 94104.

Müller, M. (2015). Assemblages and actor-networks: Rethinking sociomaterial power, politics and space. Geography Compass, 9(1), 27–41. https://doi.org/10.1111/gec3.12192

Norström, A. Vet al. (2020). Principles for knowledge co-production in sustainability research. Nature Sustainability, 3(3), 182–190. https://doi.org/10.1038/s41893-019-0448-2

Pudup, M. B. (2008). It takes a garden: Cultivating citizen-subjects in organized garden projects. Geoforum, 39(3), 1228–1240. https://doi.org/10.1016/j.geoforum.2007.06.012

Puig de la Bellacasa, M. (2017). Matters of Care: Speculative Ethics in More Than Human Worlds. University of Minnesota Press.

Rigolon, A. (2016). A complex landscape of inequity in access to urban parks: A literature review. Landscape and Urban Planning, 153, 160–169. https://doi.org/10.1016/j.landurbplan.2016.05.017

Rosol, M. (2017). Gemeinschaftlich gärtnern in der neoliberalen Stadt. Umkämpftes Grün. Zwischen neoliberaler Stadtentwicklung und Stadtgestaltung von unten, 11-32.

Säumel, I., Reddy, S. E., & Wachtel, T. (2019). Edible City Solutions—One step further to foster social resilience through enhanced socio-cultural ecosystem services in cities. Sustainability, 11(3), 972. https://doi.org/10.3390/su11030972

Sipos, Y., Battisti, B., & Grimm, K. (2008). Achieving transformative sustainability learning: Engaging head, hands and heart. International Journal of Sustainability in Higher Education, 9(1), 68–86. https://doi.org/10.1108/14676370810842193

Star, S. L., & Bowker, G. C. (2006). How to infrastructure. In L. A. Lievrouw & S. Livingstone (Eds.), Handbook of New Media: Social Shaping and Consequences of ICTs (Updated student edition, pp. 230–245). Sage.

Suchman, L. A. (2007). Human–Machine Reconfigurations: Plans and Situated Actions (2nd ed.). Cambridge University Press.

Taliep, N., & Ismail, G. (2023). Community mapping method. In Handbook of social sciences and global public health (pp. 1-22). Cham: Springer International Publishing.

van Dooren, T., Kirksey, E., & Münster, U. (2016). Multispecies studies: Cultivating arts of attentiveness. Environmental Humanities, 8(1), 1–23. https://doi.org/10.1215/22011919-3527695

Wynne, B. (1996). May the sheep safely graze? A reflexive view of the expert–lay knowledge divide. In: S. Lash, B. Szerszynski & B. Wynne (Eds.), Risk, Environment and Modernity: Towards a New Ecology (pp. 44–83). London: Sage.

Zeidler E., Kahn D.(2014). It's Debatable: Using Socioscientific Issues to Develop Scientific Literacy, K-12. NSTA Press, National Science Teachers Association. ISBN: 978-1-938-94600-4.

Responsible Agri-Food Research: A Behavioural Perspective

Madita Amoneit

Free University of Berlin, Germany

Food4Future (F4F), Leibniz Institute of Vegetable and Ornamental Crops, Germany

DOI 10.3217/978-3-99161-062-5-014, CC BY 4.0

Abstract. The agri-food system experiences pressures for a socially-desirable and sustainable transformation. The Responsible Research and Innovation (RRI) approach can arguably contribute towards a transition to more sustainable agri-food systems. However, its successful implementation in the agri-food context remains challenging. This study examines if and how agri-food researchers enact the RRI principles – particularly inclusion and anticipation – and identifies influencing factors at the individual level. Findings indicate that inclusive behaviours, such as stakeholder engagement, are more common than anticipatory behaviours. A cluster analysis reveals two behavioural patterns: 'Anticipatory Collaborators' and 'Non-anticipative Collaborators' both engaging stakeholders in their agri-food research but the latter show less anticipatory behaviours. Supporting agri-food researchers in improving their skills and creating conducive organisational environments could enhance their engagement in responsible research behaviours. By introducing a behavioural lens to RRI, this study enhances the understanding of its enactment and underscores the role of individual researchers in advancing a responsible agri-food transition.

1 Introduction

The agri-food sector encompasses challenges in economic, environmental, and societal dimensions (e.g., climate change, population growth, reduction in arable land) (Bodirsky et al., 2020; Fedoroff, 2015; Food and Agriculture Organization of the United Nations, 2023; Leclère et al., 2020). The sector is under pressure for radical transformations. Hereby, it is important that the profound transformations in the agri-food sector are implemented in a socially-desirable way by considering societal needs and by contributing to solving current challenges without creating new ones.

Following the Responsible Research and Innovation (RRI) principles was argued to contribute towards a responsible digital 'Agri-food 4.0' transition (in reference to 'Industry 4.0') (Klerkx & Rose, 2020) and enhance the positive impacts while proactively addressing emerging challenges (Rose et al., 2021). In the author's view, RRI could support not only the digital transformation but also other aspects of agri-food system transformation. Moreover, in context of system change, the need for innovative and collaborative solutions to ensure robust and resilient agri-food systems in the future is evident (Herrero et al., 2020; Lezoche et al., 2020; Preiss et al., 2022), emphasising the need to incorporate the four guiding principles of the RRI framework in the agri-food sector (Castilla-Polo & Sánchez-Hernández, 2022; Mangelkramer, 2024).

Therefore, the study focuses on the enactment of RRI principles in agri-food research practice, with a particular focus on inclusion and anticipation. Recognising that the enactment of RRI principles remains challenging, the study examines how researchers' skills, motivation, and organisational environment enable or hinder researchers' engagement in responsible behaviours. Accordingly, it addresses the following research question: *Whether and how are the principles of inclusion and anticipation enacted in agri-food research, and what factors enable or inhibit their enactment?* A survey among agri-food researchers in Germany was conducted to answer this question. The study emphasises the central role of individual researchers in driving a responsible agri-food transition. Its novelty lies in applying a behavioural lens to RRI by using the COM-B behavioural model to explore the behavioural dimensions underlying the enactment of RRI. Since this is the first attempt to take a behavioural perspective on RRI in the agri-food context, the research follows an exploratory approach, providing a foundation for future studies.

The paper is structured as follows: Section 2 introduces the RRI framework in the agri-food context and reviews current literature on the implementation of the inclusion and anticipation principles, identifying key drivers and introducing a behavioural lens to RRI. Section 3 outlines the methodology, while Section 4 presents the findings, which are discussed along with the study's limitations in Section 5. Section 6 concludes the paper.

2 Theoretical Background

RRI is a process-orientated framework developed to ensure that research and innovation are conducted in an ethical and socially-responsible manner (Owen et al., 2013; von Schomberg, 2011; Stilgoe et al., 2013). It involves engaging with society to guide the research and innovation processes and considering their broader impacts (Owen et al., 2013). RRI is defined as 'a transparent, interactive process by which societal actors and innovators become mutually responsive to each other with a view to the (ethical) acceptability, sustainability and societal desirability of the innovation process and its

marketable products […]' (von Schomberg, 2011, p. 50). It incorporates research and innovation practices that are aligned with four guiding principles: (1) *inclusion* which involves actively engaging with multiple stakeholders to ensure that a variety of values, needs, and concerns are taken into account; (2) *anticipation* which involves considering the potential short and long-term consequences of research and innovation, both positive and negative, with the aim to mitigate negative impacts; (3) *reflexivity* which refers to reflecting on own values, interests, and potential biases and how they might affect the alignment of the research and innovation with societal values and ethical considerations; and (4) *responsiveness* which involves making active efforts to respond to the insights gained through inclusion, anticipation, and reflexivity by adapting the research and innovation trajectory accordingly (Owen et al., 2013; Stilgoe et al., 2013).

It is argued that RRI has the potential to drive just and sustainable transformative change (Purvis et al., 2023; Rose et al., 2021) by broadening the scope of the 'techno-centric' agri-food sector (Jakku et al., 2023; Psarikidou, 2023). Thus, RRI can play a crucial role in facilitating a responsible agri-food transition (Klerkx & Rose, 2020).

There is substantial literature on RRI in agri-food. The review by Sabio and Lehoux (2024) reveals that the interpretation and application of RRI varies widely, although the majority of reviewed articles based on von Schomberg's (2011) definition. The studies evaluate the significance of individual RRI principles differently where some authors highlight the principle of including multiple stakeholders, others set their emphasis on anticipation of potential impacts (Sabio & Lehoux, 2024).

This study examines the inclusion and anticipation principles, in line with previous research that has adopted a selective focus (Fleming et al., 2021; Sabio & Lehoux, 2024). While the four RRI principles are closely interconnected and ideally addressed as a whole, a targeted investigation allows for a more in-depth and methodologically coherent analysis. Specifically, inclusion and anticipation are the most directly operationalisable within the chosen explorative, quantitative study design. The intention is not to diminish the importance of the reflexivity and responsiveness principles, which remain central to RRI. Reflexivity and responsiveness are often more effectively examined as subsequent or complementary dimensions that build upon prior inclusive and anticipatory activities. For instance, responsiveness can steer research and innovation towards desired trajectories, especially when grounded in inclusive, anticipatory, and reflexive considerations (Owen et al., 2013; Stilgoe et al., 2013). Concentrating on the latter two principles therefore strengthens the methodological validity and analytical depth of this study.

2.1 Inclusion in Agri-Food Research

Looking at negative examples in the agri-food sector, the importance of inclusion immediately becomes apparent. The involvement of farmers in seed innovations in Canada declined due to the prioritisation of high-yield harvests, the shift to lab-based plant breeding, seed commercialisation, and the rising influence of agri-food industry (Bronson, 2015). This exclusion is problematic as it has led to farmer dependence on chemicals, the loss of traditional agricultural knowledge, environmental harm including reduced crop diversity, and food safety concerns (Bronson, 2015). However, inclusion is often neglected in the quite 'techno-scientific' focused agri-food sector (Psarikidou, 2023). Hence, certain stakeholder groups and their expertise are excluded, despite being essential for overcoming hierarchical knowledge production in the bio-economy (Psarikidou, 2023) and ensuring a more diverse and inclusive range of perspectives. For instance, inclusion has been shown to be crucial for re-evaluating plant breeding and seed systems to address the rapid changes and ensure a sustainable and resilient future (Lopes, 2023), for identifying stakeholder needs in the development of a digital platform in the sweet potato industry (Grieger et al., 2022), for broadening discussions on agricultural robotics to enhance reflections on sustainability and justice of food production systems (Ayris et al., 2024). Inclusion helps to understand stakeholder perceptions and challenges in shaping responsible nanotechnology in the agri-food sector (Grieger et al., 2021), strengthen responsible protein transitions (Amoneit et al., 2024), and ensure economically viable, environmentally sustainable, and socially-desirable solutions in precision agriculture (Gardezi et al., 2024; Gardezi et al., 2022). An effective collaboration among diverse stakeholders through inclusive and multidisciplinary dialogue can also shed light on potential tensions within and among stakeholder groups in dairy farming (Henchion et al., 2022).

However, caution is needed, as inclusion is often narrowly interpreted as multidisciplinary research or assumed to be achieved simply by involving end-users (Jakku et al., 2022). Inclusion goes beyond knowledge exchange between scientific disciplines and especially involve stakeholder groups hard to reach (Rose et al., 2023).

The question therefore arises how inclusion can be facilitated in research and innovation processes in order to contribute to developing socially-acceptable, sustainable, and effective agri-food technologies and innovations (Henchion et al., 2022; Lopes, 2023) while fostering a joint understanding of challenges and needs to find most promising solutions for the agri-food sector (Jakku et al., 2022). Previous studies have addressed this issue by focusing either on the system-level and policy measures to make agri-food research more responsible (Klerkx & Rose, 2020; Lopes, 2023; Regan, 2019) or on the organisation and structure of research networks and projects (Jakku et al., 2022; Psarikidou, 2023; Regan, 2021). Some authors investigated different levels covering individual researchers, organisational structures (e.g., research programmes and

projects), and the socio-political context including policy measures (Jakku et al., 2023; Kuzma, 2022; Merck et al., 2022). Besides the emphasised need for structural and institutional changes to enhance stakeholder engagement in agri-food research (Jakku et al., 2022; Psarikidou, 2023; Regan, 2021), enabling individual researchers is also considered (Grieger et al., 2021; Jakku et al., 2022; Jakku et al., 2023; Kuzma, 2022; Kuzma & Cummings, 2021; Regan, 2021).

2.2 Anticipation in Agri-Food Research

The enactment of the principle of anticipation entails that researchers and innovators identify potential future impacts of their research and innovation activities before such technologies are brought into use and diffused widely (Regan, 2019; Strand et al., 2022). Anticipatory practices have been shown to deepen the understanding of possible consequences of using insects as salmon feed, including concerns about food and feed safety, fish health, pollution and waste efficiency, allowing ethical and environmental considerations to be integrated early in the research process (Strand et al., 2022). Additionally, recognising both the positive (e.g., improvements in decision-making through data availability) and negative (e.g., concerns about data sharing and ethics) impacts of introducing smart farming technologies, can help to address societal concerns at an early stage and exemplified the adoption of an RRI approach (Regan, 2019). Anticipating different future options for digital agriculture can aid to identify opportunities for improved decision-making and the consequences of different transition pathways while underscoring the need for collaboration among researchers and policymakers to shape more desirable outcomes for the digital future of agriculture (Fleming et al., 2021). Anticipation contributed towards a comprehensive view of ‘Agriculture 4.0’ and its potential impacts on livestock farming (Eastwood et al., 2021). In alignment, potential impacts should be taken into account in their embedded agri-food system instead of anticipating consequences isolated for each agri-food innovation (Klerkx & Rose, 2020). Anticipating potential positive and negative impacts meant to be incorporated along the whole research and innovation processes to contribute to responsible agri-food transitions (Klerkx & Rose, 2020) and to socially-desirable and sustainable agri-food transition pathways in the future (Mangelkramer, 2024).

However, anticipating potential impacts of digital transformation in agriculture is often limited to risk assessment and impact identification with minimal consideration of unintended consequences or broader stakeholder effects (Jakku et al., 2022). Although concerns such as data security are recognised, the focus remains on positive outcomes, neglecting potential challenges posed by future uncertainties, such as regulatory changes (Jakku et al., 2022). Therefore, special attention needs to be taken to anticipatory behaviours encompassing impacts at various scales including all potential affected stakeholders even if it might be difficult (Rose & Chilvers, 2018). However, responsible innovation means to stakeholders in the field of nanotechnology in agri-food

research considering environmental, health, and safety impacts as well as increasing product efficacy and efficiency (Grieger et al., 2021; Kokotovich et al., 2021), whereas project leaders in digital agri-food projects refer rather to management tasks (Jakku et al., 2022).

Consequently, there is a lack of understanding on how to facilitate anticipatory activities on the individual researcher level. Several studies focus on identifying and assessing potential positive and negative impacts of agri-food research applying various methods (e.g., foresight workshop, interviews, sociotechnical imaginaries) (Fleming et al., 2021; Jakku et al., 2022; Regan, 2019; Strand et al., 2022). Some authors argue that an inclusive and collaborative environment is required to engage in anticipatory activities (Jakku et al., 2022; Klerkx & Rose, 2020). Moreover, envisaged positive impacts are more inclined to be achieved and negative impacts of the research are more likely to be reduced when agri-food research is better aligned with societal values and needs (Jakku et al., 2022). However, the role of individual skills, motivators, and organisational factors is only given little attention in the literature.

2.3 Drivers of Responsible Agri-Food Research

Agri-food researchers require specific skills (e.g., systems thinking, communication) and training in order to research and innovate responsibly in agri-food (Cummings et al., 2021; Jakku et al., 2022). Scholars mostly refer to methods and tools (e.g., design thinking methods, vale-sensitive design) researchers should apply (Jakku et al., 2022; Jakku et al., 2023) whereas others highlight the role of social scientists being primarily in charge as experts to (better) align research and innovation with societal values and needs (Jakku et al., 2022).

Incentives and rewards can increase researchers' motivations towards more responsible agri-food research (Jakku et al., 2023; Kuzma, 2022; Merck et al., 2022). Drivers to pursue responsible innovation in nanotechnology encompass a range of societal, environmental, ethical, and industry-related considerations which not reflect the breadth of RRI (Kokotovich et al., 2021). In some occasions, RRI resonates with the researchers' academic values highlighting its alignment with their disciplines' missions, such as sustainability in environmental engineering or research integrity and ethics (Kokotovich et al., 2021). Some agri-food researchers emphasise that stakeholder engagement is part of their professional role and responsibility (Kokotovich et al., 2021) whereas others believe that basic research is not suitable for enacting RRI (Roberts et al., 2020).

It is suggested that agri-food researchers regularly discuss the meaning of responsible innovation creating 'an opportunity to reflect upon their own research and innovation in a broader societal context' (Grieger et al., 2021, p. 10). Reflecting on the research individually or with others can aid to consider potential research impacts (Jakku et al., 2022). Guidance provided by the organisation (e.g., code of conduct) and evaluation

systems (Jakku et al., 2023; Merck et al., 2022) can also facilitate conducting agri-food research responsibly. Funding agencies can promote responsible agri-food research by providing sufficient resources including time (Kuzma, 2022; Regan, 2021), designing 'funding and project management' more flexible (Jakku et al., 2023; Roberts et al., 2020), and setting requirements for anticipatory activities (Merck et al., 2022). Institutional barriers need to be reduced and a supportive organisational environment (e.g., 'formalised mechanisms for anticipation exercises') should be established to facilitate responsible agri-food research (Jakku et al., 2022; Jakku et al., 2023; Kuzma, 2022; Regan, 2021).

In summary, facilitating the RRI principles inclusion and anticipation can be achieved through multiple individual levers in order to support responsible agri-food research. Researchers' skills, motivational aspects, and organisational and funding environment play an important role. A better understanding of the role of individual agri-food researchers is needed to increase the enactment of inclusion and anticipation. This can help to shed light on the behavioural dimension of RRI – which appears as a current 'black box' – and can contribute to responsible agri-food research.

2.4 A Behavioural Perspective on Inclusion and Anticipation

The study takes a behavioural perspective on RRI and examines whether and how inclusion and anticipation are enacted in agri-food research by assessing the types of stakeholder groups engaged, impacts of their research and innovation anticipated, and frequency of such behaviours. Potential influencing factors are investigated by applying the behavioural COM-B model by Michie et al. (2011). It helps to dive deeper into what enables and hinders agri-food researchers to show inclusive and anticipatory behaviours. While numerous behavioural models exist that might be suitable to apply to RRI behaviours, none have yet been linked to the concept of RRI. However, a behavioural model was sought that (a) is not specialised for particular fields of applications or disciplines (e.g., HAPA model by Schwarzer (1992)) and (b) considers internal and external factors that influence behaviour. On that basis, it was decided to proceed with the behavioural change COM-B model which is applicable to behaviours across all domains and at various levels ranging from individuals, groups to entire populations (Michie et al., 2014). It enjoys a wide range of application (e.g., researchers' publishing behaviours (Weckowska et al., 2017), hand hygiene behaviours (Lambe et al., 2020)). The COM-B model consists of three components, namely capability, opportunity, and motivation, that lead to the target behaviour (Michie et al., 2005; Michie et al., 2011). *Capability* is defined as the 'individual's psychological and physical capacity to engage in the activity concerned. It includes having the necessary knowledge and skills.' (Michie et al., 2011, p. 4). *Opportunity* covers 'all the factors that lie outside the individual that make the behaviour possible or prompt it' (Michie et al., 2011, p. 4) whereas *motivation* includes 'all those brain processes that energize and direct behaviour, not just goals and

conscious decision-making' (Michie et al., 2011, p. 4). The COM-B model helps to better understand the target behaviour and its determinants while considering behaviour as part of a system related to other behaviours, not occurring in isolation (Michie et al., 2014; Michie et al., 2011). Hence, the behavioural COM-B model provide a valuable lens for examining researchers' inclusive and anticipatory behaviours and their underlying influences.

3 Methods

The study is part of the research project food4future, funded by the former German Federal Ministry of Education and Research's funding line 'Agricultural Systems of the Future' (Grant number: 031B0730H). It is based on an online survey conducted between April, 11 2022 and April, 13 2023. The results presented in this paper are drawn directly from this survey, which targeted agri-food researchers in Germany. The survey was distributed via email invitations to researchers in the food4future project as well as to researchers in similar research fields, who were identified through a comprehensive web search.

3.1 Sample

A total of 41 participants fully completed the survey and are included in the data analysis. The researchers are primarily from the fields of natural sciences (51.2%), agricultural sciences (22.0%), and social sciences (12.2%). 20 participants identify as male, 19 as female and two did not disclose their gender. In terms of career stage, 14 participants (34.1%) indicate being fully independent researchers, twelve participants being PhD researchers (29.3%), ten participants being mid-career researchers (24.4%), four participants being early-career researchers (9.8%), and one participant did not indicate their career stage (2.4%).

3.2 Measurements

The measurement of whether and how inclusion and anticipation are enacted by agri-food researchers was guided by van de Poel's (2020) suggested two-step procedure for operationalising RRI, which was originally aimed to assess RRI performance. This approach was considered appropriate for the present study, as it addresses the same methodological challenge of lacking operationalisation and available measurements. Similar to the aims of operationalising RRI performance (van de Poel, 2020) or moral values (Kroes & van de Poel, 2015), this study strived to operationalise inclusive and anticipatory behaviours and their influencing factors to make them measurable. The first step was to identify the key dimensions of inclusion and anticipation which may not be directly measurable. Second, these dimensions were translated into measurable items

which served as proxies to assess whether and how inclusion and anticipation are enacted in agri-food research. It was differentiated between anticipation of environmental and social impacts in line with previous studies (Grieger et al., 2021; Kuzma & Cummings, 2021). The focus laid both on assessing the frequency of inclusive and anticipatory behaviours during a twelve-month period and on specific characteristics of inclusive and anticipatory behaviours (e.g., stakeholder groups engaged, types of environmental and social impacts considered).

As said before, the behavioural COM-B model by Michie et al. (2011) was applied. Each of the three components – capability, opportunity, and motivation – can be subdivided into sub-components (Michie et al., 2014; Michie et al., 2011) and expanded into 14 domains by using the Theoretical Domains Framework (TDF) (Cane et al., 2012; Michie et al., 2014). Following recommendations to select domains relevant to a specific behavioural context, the focus laid on seven TDF domains, which were perceived having the potential to explain RRI behaviours. The items developed surveying the influencing factors were based on previous studies using the COM-B model and its TDF extension (Cane et al., 2012; Huijg et al., 2014; Keyworth et al., 2020; Michie et al., 2014) (see Table 1).

COM-B Component	TDF domain	Items[1]
Capability	Skills	I have the skills needed to engage non-academic stakeholders in my research.
		I have the skills needed to anticipate the environmental/ social impacts of agri-food innovations.
Opportunity	Social influences	My colleagues want me to engage non-academic stakeholders in my research.
		My colleagues want me to anticipate the environmental/ social impacts of agri-food innovations.
	Environmental context and resources	With my current workload, I have enough time to engage non-academic stakeholders in my research.
		With my current workload, I have enough time to anticipate the environmental/ social impacts of agri-food innovations.
Motivation	Professional role	Engaging non-academic stakeholders in research is part of my professional role.
		Anticipating the environmental/ social impacts of agri-food innovations is part of my professional role.
	Optimism	I am enthusiastic about engaging non-academic stakeholders in my research.
		I am enthusiastic about anticipating the environmental/ social impacts of agri-food innovations.
	Beliefs about consequences	Engagement of non-academic stakeholders in research helps to build sustainable agri-food systems of the future.
		I believe that anticipating the environmental/ social impacts of agri-food innovations helps to build sustainable future food systems.
	Intentions	In the next twelve months, I intend to engage non-academic stakeholders in my research.
		In the next twelve months, I intend to anticipate the environmental/ social impacts of agri-food innovations.

Table 1. COM-B items for Responsible Research and Innovation (RRI) behaviours (inclusion, anticipation)

[1] Items were measured on self-assessment basis on a six-point Likert-scale (strongly agree to strongly disagree). The environmental and social impacts were surveyed separately and are presented together here for reasons of readability.

3.3 Procedure

Two versions of the survey were used: one version for researchers within the food4future project and one for researchers in similar fields outside the project. The only difference was that 'food4future sub-project' replaced 'your selected research project'. Both questionnaires began with participant information, consent, and data protection, followed by questions on demographic (gender), disciplinary background, and career stage. Followed by three blocks with questions addressing researchers' behaviours: (1) engaging non-academic stakeholders (inclusion), (2) anticipating environmental impacts, and (3) anticipating social impacts. Each block had the same structure, including questions on behaviours, their frequency, and COM-factors influencing the corresponding RRI behaviour.

3.4 Data Analysis

The study applied the COM-B model to individual RRI behaviours using an exploratory approach. Data were analysed in IBM SPSS Statistics (Version 28) (IBM Corp.). 36 cases were excluded due to missing values or lack of consent. Given the ordinal nature and skewed distribution of the Likert-scaled data, median (middle value in a set of ordered data) and mode (most frequently occurring value in data) were used for descriptive statistics, as the mean value may misrepresent the 'central tendency' (Jamieson, 2004; Sullivan & Artino, 2013). In line, non-parametric statistical methods were employed, as the data did not meet normality assumptions, supporting this analytical choice despite the debate over the use of parametric tests for Likert-scaled data (Carifio & Perla, 2008; Norman, 2010; Sullivan & Artino, 2013).

Additionally, a cluster analysis was conducted to explore behavioural patterns. Following recommendations for ordinal data, Ward's method with squared Euclidean distance was applied in a hierarchical cluster analysis to determine the number of clusters (Žiberna et al., 2004). In a second step, the k-means algorithm was employed to refine and further explore the cluster structure identified in the first step.

4 Results

In the following, the study's results are presented structured by whether (frequency assessments) and how (which stakeholder groups are engaged and impacts are anticipated) inclusion and anticipation are enacted by agri-food researchers followed by the findings of the cluster analysis. Supplementary information, including sample characteristics, frequencies of study variables, dendrogram from the cluster analysis, and results of post-hoc Mann-Whitney U tests, is available from the author upon request.

4.1 Responsible Agri-Food Research

Non-academic stakeholders were engaged (inclusion) at a median level of Mdn = 3.00 (three or four times in a twelve-month period), with a mode of Mode = 2.00 (once or twice) (multiple modes exist, the smallest value is reported). Environmental impacts were considered (anticipation of environmental impacts) at a median frequency of Mdn = 2.00 (once or twice) and the mode was Mode = 1.00 (never). The anticipation of social impacts had a median of Mdn = 1.00 (never) and a mode of Mode = 1.00 (never). The findings indicated that researchers in agri-food research most frequently involved non-academic stakeholders in their research, less frequently anticipated environmental impacts and rarely anticipated social impacts. **Fig. 1** displays the frequency distribution. To obtain an overall picture of inclusive and anticipatory behaviours, specific characteristics of each behaviour were assessed, as presented in the following.

Inclusion. The researchers primarily involved 'established commercial companies' (n=20, 48.8%), followed by 'early adopters of innovation (e.g., consumers, users)' (n=15, 36.6%), and 'government agencies' (n=14, 34.1%) (multiple answers were possible). Potential future adopters (n=12, 29.3%) and civil society organisations (n=11, 26.8%) were included less frequently. Five researchers (12.2%) indicated that they did not engage any non-academic stakeholders in the last twelve months and were thus not included in the further analysis. Non-academic stakeholders were involved in particular by being informed about the research topic and the research process (n=27, 75.0%), by providing information needed for their research (n=19, 52.8%), and by giving feedback on the research process or the planned innovation (n=13, 36.1%) (multiple answers are possible).

Anticipation of environmental impacts. The majority of researchers (n=25, 61.0%) stated that they considered environmental impacts in their research during the last twelve months whereby 16 researchers (39.0%) indicated that they did not investigate any environmental impacts. Among the researchers who have considered environmental impacts, water consumption (n=17, 68.0%), land use (n=17, 68.0%), greenhouse gases emissions (n=13, 52.0%), and energy consumption (n=13, 52.0%) were the most frequently investigated environmental impacts. The most frequent practices to anticipate

environmental impacts were that the researchers clearly identified the environmental problems which can be addressed by agri-food innovations (n=16, 64.0%), they conducted pilot studies to evaluate different environmental impact scenarios (n=9, 36.0%), and technology assessment (e.g., cost-benefit analysis, life cycle analysis, etc.) (n=9, 36.0%).

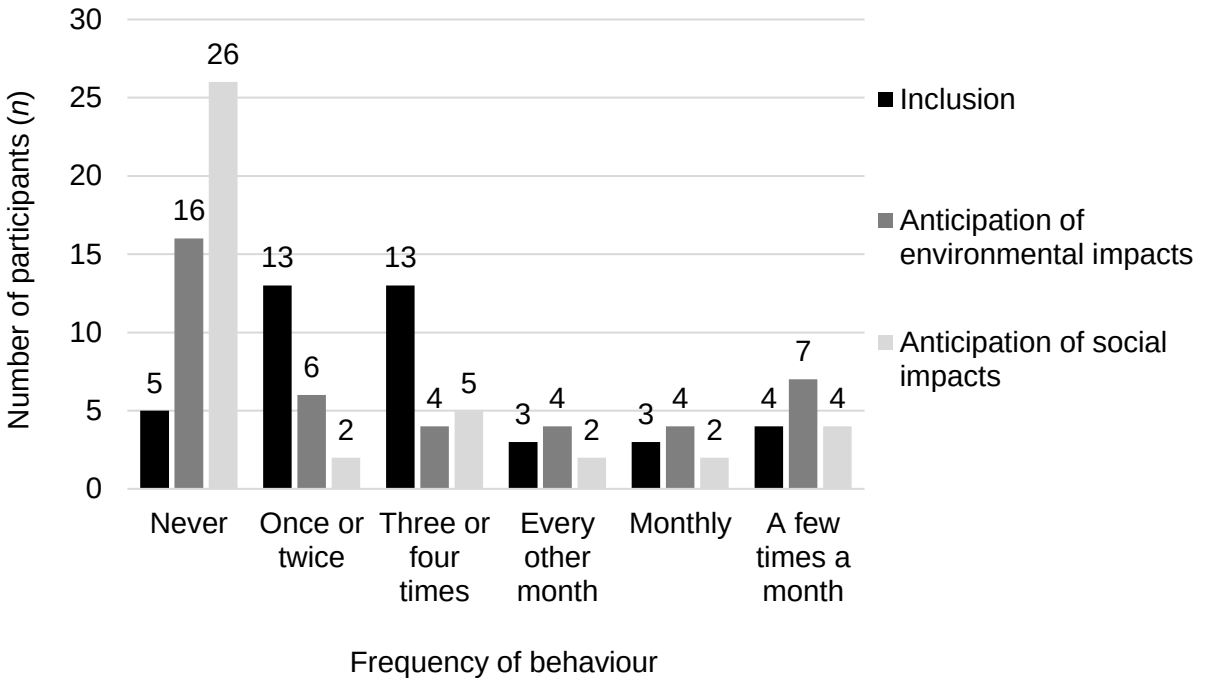


Figure 5.: Frequency of RRI behaviours (inclusion, anticipation) referring to the last twelve months.

Anticipation of social impacts. The majority of researchers (n=26, 63.4%) indicated that they did not investigate any social impacts in their research projects during the last twelve months. 15 participants (36.6%) stated that they considered social impacts – most frequently impacts on consumers (n=11, 73.3%), especially consumers' health and safety (n=8), followed by impacts on workers in the value chain (n=9, 60.0%), especially the creation and elimination of employment opportunities in the agri-food system (n=6). The most frequently practices to anticipate social impacts were that they clearly identified the social needs which can be addressed by agri-food innovations (n=10, 66.7%), employing exercise in which they tried to imagine the worst-case scenario of misuse/ misemployment/ evil use of agri-food innovations to explore potential risks (n=5, 33.3%), and they conducted technology assessment (e.g., social life cycle analysis) (n=4, 26.7%).

4.2 Drivers of Inclusion and Anticipation

A two-step cluster analysis was conducted to identify behavioural patterns of agri-food researchers along the frequency of their inclusive and anticipatory behaviours. The hierarchical cluster analysis using Ward's method and squared Euclidean distance suggested an optimal solution of two clusters, determined based on the agglomeration

schedule and dendrogram inspection. To validate this two-cluster structure, the k-means clustering algorithm was applied and confirmed the classification. The iteration history indicated convergence after four iterations. The distance between the final cluster centres was 4.39, suggesting clear separation between the two clusters. The clusters varied in size with Cluster 1 consisting of 13 (31.7%) and Cluster 2 of 28 participants (68.3%). The post-hoc Mann-Whitney U-tests showed that the two clusters did not differ in the frequency of inclusive behaviours (U = 149.50, Z = -.943, p = -.357) but in anticipatory behaviours of environmental impacts (U = 11.00, Z = -4.967, p < .001) and social impacts (U = 42.00, Z = -4.552, p < .001), whereas for the latter only the distributions differed significantly (Kolmogorov-Smirnov p < .05) (see **Fig. 2**).

The two cluster groups were labelled ‘Anticipative Collaborators’ and ‘Non-anticipative Collaborators’, as they mainly differed in the level of anticipatory behaviours. Furthermore, the two clusters differed between disciplinary backgrounds based on the Fisher's exact test (p = .016). The ‘Anticipative Collaborators’ had disciplinary backgrounds in agricultural science (n=5, 12.5%), social sciences (n=4, 10.0%), natural sciences (n=3, 7.5%), and engineering and technologies (n=1, 2.5%) whereas the ‘Non-Anticipative Collaborators’ mostly had a natural sciences background (n=18, 45.0%), followed by agricultural sciences (n=4, 10.0%), engineering and technologies (n=2, 5.0%), and social sciences, humanities and medical sciences (n=1, 2.5% each). Differences between the two clusters in gender and career stage had not been found. Overall, the cluster analysis revealed meaningful behavioural patterns among agri-food researchers, offering insights into the differences in the frequency of their anticipatory behaviours and researchers’ disciplinary backgrounds.

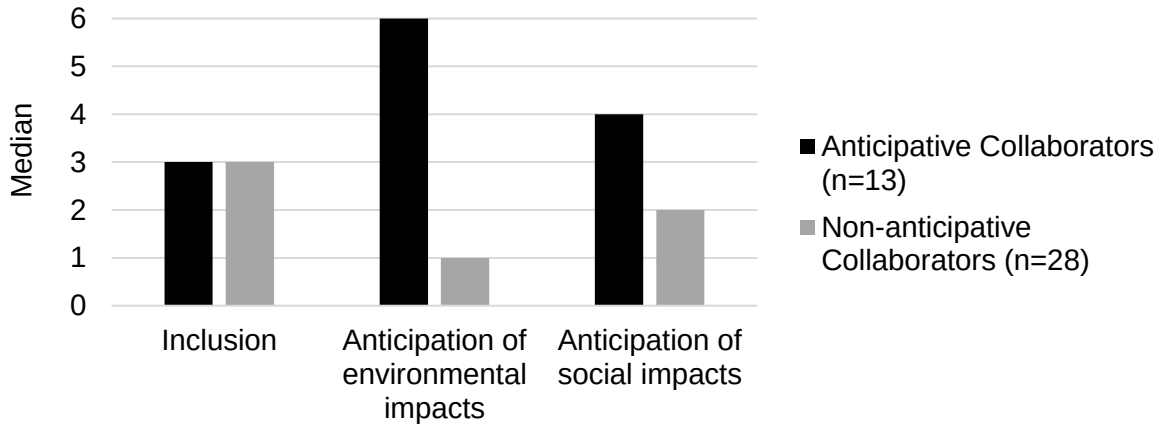


Figure 6.: Two-step cluster analysis – median differences in the frequency of inclusive and anticipatory behaviours. Note: Frequency was measured on a six-point ordinal scale and related to a twelve-month period (1 = never, 2 = once or twice, 3 = three or four times, 4 = every other month, 5 = monthly, 6 = a few times a month).

In order to investigate the influencing factors, the differences in capability, opportunity, and motivation between the two clusters were further examined. The differences were examined using post-hoc Mann-Whitney U-tests with the ‘Anticipative Collaborators’

cluster consistently exhibited higher medians. In the following only the significant results are reported.

Capability. The ‘Anticipative Collaborators’ perceived their skills to anticipate social impacts as significantly higher than the ‘Non-Anticipative Collaborators’ ($U = 81.50$, $Z = -2.907$, $p < .001$), whereas the skills for inclusion and anticipation of environmental impacts did not differ significantly.

Opportunity. The ‘Anticipative Collaborators’ reported the social pressure in their professional environment to engage stakeholders, to anticipate environmental and social impacts of their research as significantly higher than the ‘Non-Anticipative Collaborators’ (inclusion: $U = 104.00$, $Z = -2.236$, $p < .05$; anticipation of environmental impacts: $U = 94.00$, $Z = -2.518$, $p < .05$; anticipation of social impacts: $U = 63.00$, $Z = -3.418$, $p < .001$). The environmental context and resources did not play a significant role in the differences between the two clusters.

Motivation. The ‘Anticipative Collaborators’ perceived the anticipation of environmental and social impacts as part of their professional role as significantly higher than the ‘Non-Anticipative Collaborators’ (anticipation of environmental impacts: $U = 79.50$, $Z = -2.923$, $p < .001$; anticipation of social impacts: $U = 75.00$, $Z = -3.048$, $p < .001$). The ‘Anticipative Collaborators’ were significantly more optimistic to anticipate (anticipation of environmental impacts: $U = 77.50$, $Z = -3.022$, $p < .001$; anticipation of social impacts: $U = 109.50$, $Z = -2.084$, $p < .05$), and showed stronger intention to anticipate environmental as well as social impacts of their research (anticipation of environmental impacts: $U = 90.00$, $Z = -2.633$, $p < .001$; anticipation of social impacts: $U = 79.50$, $Z = -2.926$, $p < .001$). The ‘Anticipative Collaborators’ were significantly more convinced that anticipating environmental impacts contributes to sustainable agri-food innovations in the future ($U = 98.50$, $Z = -2.517$, $p < .05$). In contracts to the above results for anticipatory behaviours, the motivational aspects for inclusive behaviours showed no significant differences between the two clusters.

In summary, the ‘Anticipative Collaborators’ tended to have higher capabilities, especially in anticipating social impacts of their research and innovation, tended to have a social environment (opportunities) conducive to inclusion and anticipation, and stronger motivations for anticipatory behaviours compared to the ‘Non-Anticipative Collaborators’.

5 Discussion

The German agri-food sector faces economic, environmental, and societal challenges (Food and Agriculture Organization of the United Nations, 2023). The multifaceted transformation of the agri-food sector should arguably be socially-desirable and responsible. The four process principles of RRI (inclusion, anticipation, reflexivity, and

responsiveness) can serve as guidance in navigating the research and innovation processes (Klerkx & Rose, 2020). This study focused on inclusion and anticipation.

A survey was conducted assessing whether and how inclusion and anticipation principles are enacted in agri-food research and what factors influence the researchers' enactment in practice, based on the behavioural COM-B model. The study's findings showed that inclusive behaviours such as engaging stakeholders in research and innovation processes were more frequent than anticipatory behaviours among agri-food researchers in Germany. The growing expectations for inter- and transdisciplinary research – particularly in funding announcements – might lead to increased stakeholder engagement (Owen et al., 2021; van Rijnsoever & Hessels, 2011). The strong involvement of stakeholder groups such as commercial companies and early adopters in this study may also suggest that commercialisation efforts play a key role in driving stakeholder engagement rather than the ambition to understand society's needs and values to adapt the research and innovation activities in a responsive manner. Regarding anticipation of environmental and social impacts, environmental impacts of research and innovation were considered more frequently than potential social impacts which is in line with previous studies that environmental, health, and safety considerations are more often associated with responsible innovation, than societal concerns (Grieger et al., 2021).

5.1 Drivers of Responsible Agri-Food Research

The results of the two-step cluster analysis revealed two behavioural patterns: the 'Anticipative Collaborators' and the 'Non-anticipative Collaborators'. Both groups of agri-food researchers engaged with non-academic stakeholders but differed in considering environmental and social impacts of their research and innovation. The 'Anticipative Collaborators' tended to have higher *capabilities*, especially in anticipating social impacts of their research. This finding is in line with previous studies observing that socio-ethical considerations are neglected in anticipatory activities in dairy farming (Eastwood et al., 2019), which could be due to a lack of specialised skills required for enacting RRI principles (Cummings et al., 2021; Jakku et al., 2022; Jakku et al., 2023). However, educational and training programmes could foster the acquisition of necessary skills and thus facilitate the enactment of RRI (Merck et al., 2022).

The 'Anticipative Collaborators' tended to have greater perceived social *opportunities*. In line with previous research, the pressure to take responsibility into account in food system research is identified as one of four key drivers for integrating RRI (Sabio & Lehoux, 2024). Cultural constraints and discipline-specific resistance can serve as barriers to stakeholder engagement in the field of digital agriculture and synthetic biology (Regan, 2021; Roberts et al., 2020). Therefore, it requires support to develop norms aligned with RRI within academia that foster researchers' willingness and commitment to RRI behaviours (Regan, 2021). Although the study did not find significant differences in

anticipatory behaviours related to environmental context, organisational factors including lack of time and resources were perceived as important barriers to implement RRI in the literature (Ayris et al., 2024; Regan, 2021; Roberts et al., 2020; Taylor et al., 2023). A supportive environment (social and physical) is needed for shaping the agri-food transition responsibly (Jakku et al., 2023).

The 'Anticipative Collaborators' tended to have stronger *motivations* for anticipatory behaviours. In order to encourage agri-food researchers in general to anticipate potential environmental and social impacts of their research and innovation, incentives and reward systems could be implemented (Jakku et al., 2023; Kuzma, 2022; Merck et al., 2022). Motivating researchers to take new roles and responsibilities to conduct agri-food research responsibly might be helpful (Regan, 2021) whereas some researchers already perceive RRI in line with their disciplines' mission (Kokotovich et al., 2021).

According to the study's findings, important levers seem to lie in interpersonal dimensions including social influence, norms and organisational culture, and motivational dimensions. Differences in disciplinary backgrounds should also be taken into account. However, social scientists should not bear sole responsibility for considering social impacts of agri-food research as inter- and transdisciplinary research highlights the risk of researchers' multiple roles and possible tensions between roles (Bulten et al., 2021; Wittmayer & Schäpke, 2014). Instead, all agri-food researchers should be empowered to contribute, ensuring a responsible agri-food transition.

5.2 Limitations and Future Research

Some limitations of the study should be noted. First, by focusing on inclusion and anticipation, not all four RRI principles were investigated. This focus was based on the assumption that inclusion can provide the foundation of anticipatory consideration (Rose & Chilvers, 2018) and that critical self-reflection and responsive re-orientation of research and innovation are likely to be more impactful when inclusive and anticipatory activities take place as a first step. Nevertheless, future research is required to take a more holistic approach by investigating the enactment and its influencing factors of all four principles. Second, the study's findings should be viewed with caution, as this first attempt to apply a behavioural lens to RRI followed an exploratory approach with a small sample size of 41 participants, which may limit the robustness of the results. Nonetheless, the study provides important initial empirical findings into the enactment of RRI principles in agri-food research, which warrant further research. Third, the frequency of individual enactment of RRI was assessed over a period of twelve months, whereas, as noted by Repo and Matschoss (2019), RRI adoption (e.g., citizen participation) ideally occurs throughout the entire research and innovation process. Future research is needed to investigate RRI behaviours beyond the twelve-month period. Fourth, post-hoc tests after cluster analysis are used (Gere, 2023), but are not necessarily typical as the clusters to be compared are assigned rather randomly. For this reason, the post-hoc tests' findings

should be treated carefully. Future research might use further statistical methods to better understand the relationship between RRI behaviours and its influencing factors (capability, opportunity, and motivation). Notably, the COM-B model and TDF can be investigated with both quantitative and qualitative methods (Cane et al., 2012; Lambe et al., 2020). Fifth, the study surveyed agri-food researchers in Germany which limits its generalisability. Therefore, future research might consider other fields of application and further stakeholder groups across different countries.

6 Conclusions

This study makes several contributions to understanding the enactment of RRI principles in agri-food research. First, it is the first attempt to introduce a behavioural lens to RRI by applying the COM-B model to investigate factors influencing the enactment of inclusion and anticipation. Second, it provides initial empirical insights into the behavioural dimension of RRI, enhancing understanding of inclusive and anticipatory behaviours and their drivers. These insights can inform strategies to promote inclusive and anticipatory behaviours, foster reflexivity and responsiveness in research and innovation, and support socially desirable and sustainable agri-food transition pathways. Third, the study highlights the pivotal role of individual researchers as key actors in driving responsible agri-food transitions, in line with previous scholarship (Felt et al., 2018; Shelley-Egan et al., 2018).

From a practical perspective, the findings suggest that enhancing RRI enactment requires facilitating the development of relevant skills particularly for anticipating social impacts, creating social and organisational environments supportive of RRI, and strengthening individual motivation to anticipate both environmental and social impacts. As an exploratory study, it serves as a foundation for future research to further unpack researchers' responsible behaviours and the factors enabling them.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used ChatGPT-5 mini and ChatGPT-4 in order to improve readability and language of the work. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Funding sources

This research was funded by the German Federal Ministry of Education and Research, grant number 031B0730H.

Declaration of competing interest

The author declares that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

I would like to sincerely thank Dagmara Weckowska for supervision and support throughout the development of this paper.

References

Amoneit, Madita; Weckowska, Dagmara; Preiss, Myriam; Biedermann, Annette; Gellrich, Leon; Dreher, Carsten; Schreiner, Monika (2024): Public Perceptions of Alternative Protein Sources: Implications for Responsible Agrifood Transition Pathways. In Sustainability 16 (2), pp. 566. DOI: 10.3390/su16020566.

Ayris, Kirsten; Jackman, Anna; Mauchline, Alice; Rose, David C. (2024): Exploring Inclusion in UK Agricultural Robotics Development: Who, How, and Why? In Agric Hum Values 41 (3), pp. 1257–1275. DOI: 10.1007/s10460-024-10555-6.

Bodirsky, Benjamin L.; Dietrich, Jan P.; Martinelli, Eleonora; Stenstad, Antonia; Pradhan, Prajal; Gabrysch, Sabine; Mishra, Abhijeet; Weindl, Isabelle; Le Mouël, Chantal; Rolinski, Susanne; Baumstark, Lavinia; Wang, Xiaoxi; Waid, Jillian L.; Lotze-Campen, Hermann; Popp, Alexander (2020): The Ongoing Nutrition Transition Thwarts Long-term Targets for Food Security, Public Health and Environmental Protection. In Sci Rep 10 (1), pp. 19778. DOI: 10.1038/s41598-020-75213-3.

Bronson, Kelly (2015): Responsible to Whom? Seed Innovations and the Corporatization of Agriculture. In Journal of Responsible Innovation 2 (1), pp. 62–77. DOI: 10.1080/23299460.2015.1010769.

Bulten, Ellen; Hessels, Laurens K.; Hordijk, Michaela; Segrave, Andrew J. (2021): Conflicting Roles of Researchers in Sustainability Transitions: Balancing Action and Reflection. In Sustain Sci 16 (4), pp. 1269–1283. DOI: 10.1007/s11625-021-00938-7.

Cane, James; O'Connor, Denise; Michie, Susan (2012): Validation of the Theoretical Domains Framework for Use in Behaviour Change and Implementation Research. In Implementation Sci 7 (1), pp. 37. DOI: 10.1186/1748-5908-7-37.

Carifio, James; Perla, Rocco (2008): Resolving the 50-year Debate Around Using and Misusing Likert Scales. In Medical Education 42 (12), pp. 1150–1152. DOI: 10.1111/j.1365-2923.2008.03172.x.

Castilla-Polo, Francisca; Sánchez-Hernández, M. I. (2022): International Orientation: An Antecedent-consequence Model in Spanish Agri-Food Cooperatives Which are Aware of the Circular Economy. In Journal of Business Research 152, pp. 231–241. DOI: 10.1016/j.jbusres.2022.07.038.

Cummings, Christopher L.; Kuzma, Jennifer; Kokotovich, Adam; Glas, David; Grieger, Khara (2021): Barriers to Responsible Innovation of Nanotechnology Applications in Food and Agriculture: A Study of US Experts and Developers. In NanoImpact 23, pp. 100326. DOI: 10.1016/j.impact.2021.100326.

Eastwood, Callum R.; Edwards, Paul; Turner, James A. (2021): Review: Anticipating Alternative Trajectories for Responsible Agriculture 4.0 Innovation in Livestock Systems. In Animal 15 Suppl 1, pp. 100296. DOI: 10.1016/j.animal.2021.100296.

Eastwood, Callum R.; Klerkx, Laurens; Ayre, Margaret; Dela Rue, Brian T. (2019): Managing Socio-Ethical Challenges in the Development of Smart Farming: From a Fragmented to a Comprehensive Approach for Responsible Research and Innovation. In J Agric Environ Ethics 32 (5-6), pp. 741–768. DOI: 10.1007/s10806-017-9704-5.

Fedoroff, Nina V. (2015): Food in a Future of 10 Billion. In Agric & Food Secur 4 (1). DOI: 10.1186/s40066-015-0031-7.

Felt, Ulrike; Fochler, Maximilian; Sigl, Lisa (2018): IMAGINE RRI. A Card-Based Method for Reflecting on Responsibility in Life Science Research. In Journal of Responsible Innovation 5 (2), pp. 201–224. DOI: 10.1080/23299460.2018.1457402.

Fleming, Aysha; Jakku, Emma; Fielke, Simon; Taylor, Bruce M.; Lacey, Justine; Terhorst, Andrew; Stitzlein, Cara (2021): Foresighting Australian Digital Agricultural Futures: Applying Responsible Innovation Thinking to Anticipate Research and Development Impact Under Different Scenarios. In Agricultural Systems 190, pp. 103120. DOI: 10.1016/j.agsy.2021.103120.

Food and Agriculture Organization of the United Nations. (2023). *Global Food Security Challenges and its Drivers: Conflicts and Wars in Ukraine and Other Countries, Slowdowns and Downturns, and Climate Change.* FAO.

Gardezi, Maaz; Abuayyash, Halimeh; Adler, Paul R.; Alvez, Juan P.; Anjum, Rubaina; Badireddy, Appala R.; Brugler, Skye; Carcamo, Pablo; Clay, David; Dadkhah, Ali; Emery, Mary; Faulkner, Joshua W.; Joshi, Bhavna; Joshi, Deepak R.; Khan, Awais H.; Koliba, Christopher; Kumari, Sheetal; McMaine, John; Merrill, Scott, . . . Zia, Asim (2024): The Role of Living Labs in Cultivating Inclusive and Responsible Innovation in Precision Agriculture. In Agricultural Systems 216, pp. 103908. DOI: 10.1016/j.agsy.2024.103908.

Gardezi, Maaz; Adereti, Damilola T.; Stock, Ryan; Ogunyiola, Ayorinde (2022): In Pursuit of Responsible Innovation for Precision Agriculture Technologies. In Journal of Responsible Innovation 9 (2), pp. 224–247. DOI: 10.1080/23299460.2022.2071668.

Gere, Attila (2023): Recommendations for Validating Hierarchical Clustering in Consumer Sensory Projects. In Current Research in Food Science 6, pp. 100522. DOI: 10.1016/j.crfs.2023.100522.

Grieger, Khara; Merck, Ashton W.; Cuchiara, Maude; Binder, Andrew R.; Kokotovich, Adam; Cummings, Christopher L.; Kuzma, Jennifer (2021): Responsible Innovation of Nano-Agrifoods: Insights and Views from U.S. Stakeholders. In NanoImpact 24, pp. 100365. DOI: 10.1016/j.impact.2021.100365.

Grieger, Khara; Zarate, Sebastian; Barnhill-Dilling, Sarah K.; Hunt, Shelly; Jones, Daniela; Kuzma, Jennifer (2022): Fostering Responsible Innovation Through Stakeholder Engagement: Case Study of North Carolina Sweetpotato Stakeholders. In Sustainability 14 (4), pp. 2274. DOI: 10.3390/su14042274.

Henchion, Maeve M.; Regan, Áine; Beecher, Marion; MackenWalsh, Áine (2022): Developing 'Smart' Dairy Farming Responsive to Farmers and Consumer-Citizens: A Review. In Animals 12 (3), pp. 360. DOI: 10.3390/ani12030360.

Herrero, Mario; Thornton, Philip K.; Mason-D'Croz, Daniel; Palmer, Jeda; Benton, Tim G.; Bodirsky, Benjamin L.; Bogard, Jessica R.; Hall, Andrew; Lee, Bernice; Nyborg, Karine; Pradhan, Prajal; Bonnett, Graham D.; Bryan, Brett A.; Campbell, Bruce M.; Christensen, Svend; Clark, Michael; Cook, Mathew T.; Boer, Imke J. M. de; Downs, Chris, . . . West, Paul C. (2020): Innovation Can Accelerate the Transition Towards a Sustainable Food System. In Nat Food 1 (5), pp. 266–272. DOI: 10.1038/s43016-020-0074-1.

Huijg, Johanna M.; Gebhardt, Winifred A.; Crone, Mathilde R.; Dusseldorp, Elise; Presseau, Justin (2014): Discriminant Content Validity of a Theoretical Domains Framework Questionnaire for Use in Implementation Research. In Implementation Sci 9 (1), pp. 11. DOI: 10.1186/1748-5908-9-11.

IBM Corp. *IBM SPSS Statistics for Windows* (Version Version 28.0.1.0). IBM Corp.

Jakku, Emma; Fielke, Simon; Fleming, Aysha; Stitzlein, Cara (2022): Reflecting on Opportunities and Challenges Regarding Implementation of Responsible Digital Agri-Technology Innovation. In Sociologia Ruralis 62 (2), pp. 363‐388. https://onlinelibrary.wiley.com/doi/full/10.1111/soru.12366

Jakku, Emma; Fleming, Aysha; Espig, Martin; Fielke, Simon; Finlay-Smits, Susanna C.; Turner, James A. (2023): Disruption Disrupted? Reflecting on the Relationship Between Responsible Innovation and Digital Agriculture Research and Development at Multiple Levels in Australia and Aotearoa New Zealand. In Agricultural Systems 204, pp. 103555. DOI: 10.1016/j.agsy.2022.103555.

Jamieson, Susan (2004): Likert Scales: How to (Ab)use Them. In Med Educ 38 (12), pp. 1217–1218. DOI: 10.1111/j.1365-2929.2004.02012.x.

Keyworth, Chris; Epton, Tracy; Goldthorpe, Joanna; Calam, Rachel; Armitage, Christopher J. (2020): Acceptability, Reliability, and Validity of a Brief Measure of Capabilities, Opportunities, and Motivations ('COM-B'). In British Journal of Health Psychology 25 (3), pp. 474–501. DOI: 10.1111/bjhp.12417.

Klerkx, Laurens; Rose, David (2020): Dealing With the Game-Changing Technologies of Agriculture 4.0: How Do We Manage Diversity and Responsibility in Food System Transition Pathways? In Global Food Security 24, pp. 100347. DOI: 10.1016/j.gfs.2019.100347.

Kokotovich, Adam E.; Kuzma, Jennifer; Cummings, Christopher L.; Grieger, Khara (2021): Responsible Innovation Definitions, Practices, and Motivations from Nanotechnology Researchers in Food and Agriculture. In Nanoethics 15 (3), pp. 229–243. DOI: 10.1007/s11569-021-00404-9.

Kroes, Peter; van de Poel, Ibo. (2015): Design for Values and the Definition, Specification, and Operationalization of Values. In *Handbook of Ethics, Values, and Technological Design* (pp. 151–178). Springer Netherlands. DOI: 10.1007/978-94-007-6970-0_11.

Kuzma, Jennifer (2022): Implementing Responsible Research and Innovation: A Case Study of U.S. Biotechnology Oversight. In GPPG 2 (3), pp. 306–325. DOI: 10.1007/s43508-022-00046-x.

Kuzma, Jennifer; Cummings, Christopher L. (2021): Cultural Beliefs and Stakeholder Affiliation Influence Attitudes Towards Responsible Research and Innovation Among United States Stakeholders Involved in Biotechnology and Gene Editing. In Front. Polit. Sci. 3, Article 677003. DOI: 10.3389/fpos.2021.677003.

Lambe, Kathryn; Lydon, Sinéad; Madden, Caoimhe; McSharry, Jenny; Marshall, Rebecca; Boylan, Ruth; Hehir, Aoife; Byrne, Molly; Tujjar, Omar; O'Connor, Paul (2020): Understanding Hand Hygiene Behaviour in the Intensive Care Unit to Inform Interventions: An Interview Study. In BMC Health Serv Res 20 (1), pp. 353. DOI: 10.1186/s12913-020-05215-4.

Leclère, David; Obersteiner, Michael; Barrett, Mike; Butchart, Stuart H. M.; Chaudhary, Abhishek; Palma, Adriana de; DeClerck, Fabrice A. J.; Di Marco, Moreno; Doelman, Jonathan C.; Dürauer, Martina; Freeman, Robin; Harfoot, Michael; Hasegawa, Tomoko; Hellweg, Stefanie; Hilbers, Jelle P.; Hill, Samantha L. L.; Humpenöder, Florian; Jennings, Nancy; Krisztin, Tamás, . . . Young, Lucy (2020): Bending the Curve of Terrestrial Biodiversity Needs an Integrated Strategy. In Nature 585 (7826), pp. 551–556. DOI: 10.1038/s41586-020-2705-y.

Lezoche, Mario; Hernandez, Jorge E.; Del Alemany Díaz, Maria M. E.; Panetto, Hervé; Kacprzyk, Janusz (2020): Agri-food 4.0: A Survey of the Supply Chains and Technologies for the Future Agriculture. In Computers in Industry 117, pp. 103187. DOI: 10.1016/j.compind.2020.103187.

Lopes, Mauricio A. (2023): Rethinking Plant Breeding and Seed Systems in the Era of Exponential Changes. In Ciênc. Agrotec. 47, e0001R23. DOI: 10.1590/1413-70542023470001R23.

Mangelkramer, Delia (2024): Options for Making Responsive Future Strategy to Foster Sustainability Transitions in the German Agri-Food Sector: A Delphi-Based Approach. In Eur J Futures Res 12 (1), pp. 1–20. DOI: 10.1186/s40309-024-00230-8.

Merck, Ashton W.; Grieger, Khara D.; Kuzma, Jennifer (2022): How Can We Promote the Responsible Innovation of Nano-Agrifood Research? In Environmental Science & Policy 137, pp. 185–190. DOI: 10.1016/j.envsci.2022.08.027.

Michie, Susan; Atkins, Lou; West, Robert. (2014). *The Behaviour Change Wheel: A Guide to Designing Interventions* (First edition). Silverback Publishing.

Michie, Susan; Johnston, M.; Abraham, C.; Lawton, R.; Parker, D.; Walker, A. (2005): Making Psychological Theory Useful for Implementing Evidence Based Practice: A Consensus Approach. In Qual Saf Health Care 14 (1), pp. 26–33. DOI: 10.1136/qshc.2004.011155.

Michie, Susan; van Stralen, Maartje M.; West, Robert (2011): The Behaviour Change Wheel: A New Method for Characterising and Designing Behaviour Change Interventions. In Implementation Sci 6 (1), pp. 42. DOI: 10.1186/1748-5908-6-42.

Norman, Geoff (2010): Likert Scales, Levels of Measurement and the ‘Laws’ of Statistics. In Adv in Health Sci Educ 15 (5), pp. 625–632. DOI: 10.1007/s10459-010-9222-y.

Owen, Richard; Pansera, Mario; Macnaghten, Phil; Randles, Sally (2021): Organisational Institutionalisation of Responsible Innovation. In Research Policy 50 (1), pp. 104132. DOI: 10.1016/j.respol.2020.104132.

Owen, Richard; Stilgoe, Jack; Macnaghten, Phil; Gorman, Mike; Fisher, Erik; Guston, Dave (2013): A Framework for Responsible Innovation. In Responsible Innovation 1, pp. 27–50.

Preiss, Myriam; Vogt, Julia H.-M.; Dreher, Carsten; Schreiner, Monika (2022): Trends Shaping Western European Agrifood Systems of the Future. In Sustainability 14 (21), pp. 13976. DOI: 10.3390/su142113976.

Psarikidou, Katerina (2023): Configuring More Responsible Knowledge-Based Bio-Economies: The Case of Alternative Agro-Food Networks. In Journal of

Responsible Innovation 10 (1), Article 2196818, pp. 2196818. DOI: 10.1080/23299460.2023.2196818.

Regan, Áine (2019): 'Smart farming' in Ireland: A Risk Perception Study with Key Governance Actors. In NJAS - Wageningen Journal of Life Sciences 90-91 (1), pp. 1–10. DOI: 10.1016/j.njas.2019.02.003.

Regan, Áine (2021): Exploring the Readiness of Publicly Funded Researchers to Practice Responsible Research and Innovation in Digital Agriculture. In Journal of Responsible Innovation 8 (1), pp. 28–47. DOI: 10.1080/23299460.2021.1904755.

Repo, Petteri; Matschoss, Kaisa (2019): Considering Expert Takeovers in Citizen Involvement Processes. In Journal of Responsible Innovation 6 (2), pp. 119–142. DOI: 10.1080/23299460.2019.1568145.

Roberts, Pat; Herkert, Joseph; Kuzma, Jennifer (2020): Responsible Innovation in Biotechnology: Stakeholder Attitudes and Implications for Research Policy. In Elementa: Science of the Anthropocene 8, Article 47. DOI: 10.1525/elementa.446.

Rose, David C.; Barkemeyer, Anna; Boon, Auvikki de; Price, Catherine; Roche, Dannielle (2023): The Old, the New, or the Old Made New? Everyday Counter-Narratives of the So-Called Fourth Agricultural Revolution. In Agric Hum Values 40 (2), pp. 423–439. DOI: 10.1007/s10460-022-10374-7.

Rose, David C.; Chilvers, Jason (2018): Agriculture 4.0: Broadening Responsible Innovation in an Era of Smart Farming. In Front. Sustain. Food Syst. 2, Article 87. DOI: 10.3389/fsufs.2018.00087.

Rose, David C.; Wheeler, Rebecca; Winter, Michael; Lobley, Matt; Chivers, Charlotte-Anne (2021): Agriculture 4.0: Making It Work for People, Production, and the Planet. In Land Use Policy 100, pp. 104933. DOI: 10.1016/j.landusepol.2020.104933.

Sabio, Renata P.; Lehoux, Pascale (2024): Responsible Research and Innovation in Food Systems: A Critical Review of the Literature and Future Research Avenues. In Agric Hum Values, pp. 1–14. DOI: 10.1007/s10460-024-10672-2.

Schwarzer, Ralf (1992). *Self-Efficacy: Thought Control of Action: Self-Efficacy in the Adoption and Maintenance of Health Behaviors: Theoretical Approaches and a New Model*. Hemisphere. https://scholar.google.de/citations?user=w2m4eluaaaaj&hl=de&oi=sra.

Shelley-Egan, Clare; Bowman, Diana M.; Robinson, Douglas K. R. (2018): Devices of Responsibility: Over a Decade of Responsible Research and Innovation Initiatives for Nanotechnologies. In Sci Eng Ethics 24 (6), pp. 1719–1746. DOI: 10.1007/s11948-017-9978-z.

Stilgoe, Jack; Owen, Richard; Macnaghten, Phil (2013): Developing a Framework for Responsible Innovation. In Research Policy 42 (9), pp. 1568–1580. DOI: 10.1016/j.respol.2013.05.008.

Strand, R.; Gamboa, G.; Dankel, D. J.; Giampietro, M. (2022): Insect Feeds in Salmon Aquaculture: Sociotechnical Imagination and Responsible Story-Telling. In JIFF 8 (11), pp. 1205–1220. DOI: 10.3920/JIFF2020.0127.

Sullivan, Gail M.; Artino, Anthony R. (2013): Analyzing and Interpreting Data from Likert-Type Scales. In J Grad Med Educ 5 (4), pp. 541–542. DOI: 10.4300/JGME-5-4-18.

Taylor, Ken; Woods, Simon; Johns, Alex; Murray, Heath (2023): Intrinsic Responsible Innovation in a Synthetic Biology Research Project. In New Genetics and Society 42 (1), Article e2232684, e2232684. DOI: 10.1080/14636778.2023.2232684.

van de Poel, Ibo. (2020): RRI Measurement and Assessment: Some Pitfalls and a Proposed Way Forward. In *Assessment of Responsible Innovation* (pp. 339–360). Routledge. DOI: 10.4324/9780429298998-25.

van Rijnsoever, Frank J.; Hessels, Laurens K. (2011): Factors Associated With Disciplinary and Interdisciplinary Research Collaboration. In Research Policy 40 (3), pp. 463–472. DOI: 10.1016/j.respol.2010.11.001.

von Schomberg, René von. (2011): Prospects for Technology Assessment in a Framework of Responsible Research and Innovation. In *Technikfolgen abschätzen lehren: Bildungspotenziale transdisziplinärer Methoden* (1., neue Ausg, pp. 39–61). VS Verl. für Sozialwiss. DOI: 10.1007/978-3-531-93468-6_2.

Weckowska, Dagmara M.; Levin, Nadine; Leonelli, Sabina; Dupré, John; Castle, David (2017): Managing the Transition to Open Access Publishing: A Psychological Perspective. In Prometheus 35 (2), pp. 111–135. DOI: 10.1080/08109028.2017.1408289.

Wittmayer, Julia M.; Schäpke, Niko (2014): Action, Research and Participation: Roles of Researchers in Sustainability Transitions. In Sustain Sci 9 (4), pp. 483–496. DOI: 10.1007/s11625-014-0258-4.

Žiberna, Aleš; Kejžar, Nataša; Golob, Petra. (2004): A Comparison of Different Approaches to Hierarchical Clustering of Ordinal Data. In Advances in Methodology and Statistics 1 (1), pp. 57–73.

Building Research Communities through Communication: The Case of FOSSR

Serena Fabrizio, Rita Giuffredi, Alessandra Maria Stilo

Research Institute on Sustainable Economic Growth, National Research Council (CNR-IRCrES), Italy

DOI 10.3217/978-3-99161-062-5-015, CC BY 4.0

Abstract. In the age of Open Science, research communication within Open Research Infrastructures (RIs) should evolve beyond traditional dissemination models. Rather than merely transferring knowledge, communication should serve as an enabling mechanism for community-building, ensuring long-term engagement with research outputs. Open Science RIs depend on active and engaged communities to achieve their mission of collecting, curating, and sharing research data. Without sustained interaction between researchers, policymakers, stakeholder groups, and the public, infrastructures risk becoming static repositories rather than dynamic spaces for knowledge exchange.

This paper explores how research infrastructures can foster collaborative, participatory, and sustainable Open Science. We present the conceptual design and mid-term results of the communication strategy developed for the NRRP-funded FOSSR (Fostering Open Science in Social Science Research) Research Infrastructure, showing how communication can be reframed as an infrastructural function rather than an auxiliary activity.

By examining FOSSR's communication strategy, this paper contributes to ongoing discussions about the evolving role of communication in Open Science research infrastructures. It argues that research infrastructure communication must move beyond knowledge transfer to actively shaping collaborative research environments. This shift is critical for ensuring the sustainability, inclusivity, and long-term impact of Open Science initiatives.

1 Introduction

In the current era of Open Science, the way research is shared within Open Research Infrastructures (RIs) and with the public is expected to undergo a transformation that transcends conventional dissemination models. This evolution is concomitant with the discourse that has emerged over the preceding decades concerning the limitations of the deficit model of science communication (Bucchi & Neresini, 2008). The prevailing

paradigm is characterized by an increased respect for the unique knowledge contributions of individuals and, more broadly, for the characteristics of diverse audiences, fostering multidirectional exchanges and knowledge co-production (Trench, 2008).

The emergence of Open Science, proposed as a transformative framework for contemporary research, is oriented to significantly reshape the epistemic, institutional, and communicative dimensions of scientific practice. Emphasizing transparency, accessibility, and collaboration, Open Science challenges the contemporary shift of knowledge production and circulation towards privatization and compartmentalisation (Ziman, 2000), calling for subsequent coherent choices on how research is organized, shared, and evaluated (Nosek et al., 2015).

Open RIs depend on active and engaged communities to achieve their mission of collecting, curating, and sharing research data. Without sustained interaction between researchers, policymakers, stakeholders, and the public, infrastructures risk becoming static repositories rather than dynamic spaces for knowledge exchange.

A focus on engaged communities necessitates continuous attention to the sustainment and care of a group comprising a diversity of epistemic cultures (Knorr-Cetina, 1999), encompassing disciplinary and transdisciplinary actors engaged with open science and research infrastructures. Such a focus demands a sustained effort to adapt and care for the human component of the RI, including training, networking, and the opening of reflexive spaces.

In this evolving landscape, the understanding of the role and purpose of communication evolves from being a peripheral, service-like activity confined to the dissemination of results, as in most institutional or promotional communication (Nisbet & Markowitz, 2016) to become a core infrastructural and relational component of research itself, and is necessary to the building and maintenance of the engaged community.

Conventional models of scientific communication have predominantly emphasised the unidirectional transmission of information from researchers to external audiences, frequently conceptualised in terms of dissemination or outreach, and this is also true for the communication efforts of research institutions (Claessens, 2018; Trench, 2008). Nevertheless, this model is increasingly recognised as inadequate in capturing the complex communicative ecologies that characterise contemporary science-society interplay, and it is unsuitable for sustaining the complex circulation mechanisms that feed open environments. It also suffers from being grounded to clear demarcations between scientists and the public, while the community supporting a research infrastructure is hybrid and multi-actor by election, blurring clear distinctions among social groups. In these settings, communication cannot be regarded as an endpoint service to promote the dissemination of research outputs, but rather as a process that mediates collaboration, fosters knowledge exchange among the diverse layers of the community,

supports the establishment of relations, and ultimately sustains the sociotechnical systems upon which research depends.

As such, communication becomes infrastructural: it underpins the very possibility of collective knowledge production, while also reflecting broader shifts in the governance, organization, and evaluation of science (Cerroni, 2006; Davies & Horst, 2016). Rather than merely transferring knowledge, communication serves in this context as an enabling mechanism for community-building, ensuring long-term engagement with research outputs.

This paper explores how research infrastructures can foster collaborative, participatory, and sustainable Open Science by assigning a prominent role to community-building through communicative actions. We present the conceptual design and mid-term outcomes of the communication strategy developed for the NRRP-funded project FOSSR (Fostering Open Science in Social Science Research), showing how communication can be reframed as a structural function rather than an auxiliary activity.

Our central aim is to explore how communication functions as both a relational and infrastructural element within FOSSR Open Cloud. Specifically, we will describe the design of communication practices and their first operational phase at the midpoint of the project duration, highlighting how communication transforms its role to focus on mediating relationships among researchers, policymakers, and civil society actors. We will also explore how a more structural and strategic role of communication could contribute to the broader goals of Open Science, particularly in the context of social research.

2. Theoretical Framework

The development of science communication theory over the last few decades offers a critical perspective on describing the changes occurring in the field of communication in Open Science and Research Infrastructures (RIs). Traditionally, the dominant model of science communication was the ‘Public Understanding of Science’ model, which posited that public scepticism or disengagement with science stemmed from a lack, or ‘deficit,’ of knowledge on the part of citizens (Bodmer, 1985; Cortassa, 2016; Miller, 2004). According to this view, the role of communication was to transfer scientific facts from experts to lay audiences in a unidirectional flow (Nisbet & Scheufele, 2009), with features similar to those of basic school education. However, this model has been the object of scholarly criticism for its technocratic and paternalistic assumptions and its poor capacity to include the social, cultural, and political dimensions of public engagement with science (Davies & Horst, 2016; Jasanoff, 2003). In response, scholars have proposed and developed a reflection focusing on dialogic and participatory models that emphasize mutual learning, co-production of knowledge, and the legitimization and inclusion of

diverse epistemologies in both the understanding and resolution of science-based issues (Funtowicz & Ravetz, 1993; Jasanoff, 2004b; Wynne, 1992).

For the purposes of our research, we adopt the definition of Open Science as provided by UNESCO (2021), which significantly lists open science infrastructures, science communication and the engagement of societal actors among the pillars of open science: ‘an inclusive construct that combines various movements and practices aiming to make multilingual scientific knowledge openly available, accessible and reusable for everyone, to increase scientific collaborations and sharing of information for the benefits of science and society, and to open the processes of scientific knowledge creation, evaluation and communication to societal actors beyond the traditional scientific community. It comprises all scientific disciplines and aspects of scholarly practices, including basic and applied sciences, natural and social sciences and the humanities, and it builds on the following key pillars: open scientific knowledge, open science infrastructures, science communication, open engagement of societal actors and open dialogue with other knowledge systems’.

The dialogic models in science communication share a common ethos with Open Science, in that they both seek to democratize knowledge production and foster an inclusive research ecosystem (Hecker et al., 2018; Jasanoff, 2004a). In both contexts, communication is not merely about informing the public and disseminating research outputs; it is also about facilitating meaningful and long-term interactions among a diverse array of actors, including researchers, policymakers, civil society, and communities. This shift is especially relevant in the context of RIs, which function not only as technical platforms but also as social and organisational entities oriented to facilitate collaboration and knowledge exchange across institutional and disciplinary boundaries.

Within academic research, the concept of ‘productive interactions’ was introduced by Molas-Gallart & Tang (2011) is particularly relevant to capture the relational features of communication. This approach moves beyond linear models of impact assessment to focus on the interactions between researchers and stakeholders that can produce socially relevant outcomes. *Productive interactions* can be direct (e.g., face-to-face meetings), indirect (e.g., through documents or tools), or financial (e.g., funding collaborations), and they are understood as essential mechanisms through which research acquires meaning and utility beyond academia, in broader societal contexts. A key aspect of the FOSSR communication strategy's design was to consider communication practices that support such interactions as core strategic assets, rather than merely ‘ancillary’ to the infrastructure’s mission and sustainability.

When communicative practices are considered to be of such central importance to the development and long-term sustainability of RIs, the dynamics related to the shaping and interplay of communities acquire a central role. In a context centered on knowledge production, the interplay among the diverse ‘epistemic cultures’ (Knorr-Cetina, 1999), or

the specific ways in which different scientific communities construct and validate knowledge, is of notable relevance. These cultures shape not only the methods and instruments of research but also the communicative norms and expectations that govern interaction within and beyond the scientific community. The concept of communication is intricately interwoven with the epistemic fabric of research, a phenomenon that is particularly evident in the field of social science RIs, where the physical components of the infrastructure, although present and crucial to ground it, are less visible than big science laboratories. Consequently, the human components and the aspects related to the development of a robust interplay among all actors are vital. The function of communicative actions extends beyond the mere transmission of information, encompassing the negotiation of meaning, the establishment of trust, and the harmonization of diverse perspectives.

The central dimension of communication within Ris can also be highlighted by organizational and sociotechnical theories. Star and Ruhleder (1996) argued that infrastructure becomes visible only upon breakdown, stressing its embeddedness in everyday practices. Communication, in this sense, could be considered ‘infrastructural’ when it enables the seamless coordination of distributed activities, supports interoperability, and sustains the institutional ability to preserve and recollect the past within research organizations. As RIs represent structures in the research realm where the general trend towards increasing complexity, hybridization, and transnationality prevails, the design of communicative processes within RIs, ranging from digital platforms to governance protocols, becomes a critical site of sociotechnical negotiation and potential innovation.

Contemporary research communication is also assigned a central role in linking scientific practice to broader questions of citizenship, policy, and institutional legitimacy, especially concerning issues positioned at the superposition of different disciplinary and sectoral perspectives and worldviews, such as the governance of socio-ecological problems or the conditions for fostering inclusive sustainability. Scholars have emphasized the importance of public communication in shaping the social contract of science, particularly in contexts where public funding and political accountability are at stake (Benessia et al., 2016; Irwin, 2015; Jasanoff, 2003, 2005; Wynne et al., 2007). Nowotny et al. (2001) argue that science must become more socially robust by engaging with the values, expectations, and uncertainties of society. This requires communicative practices that are not only transparent but also dialogic and responsive, able to foster the exercise of the emerging right to scientific citizenship, or the right to appropriately exercise the right to the governance of scientific knowledge in the knowledge society (Cerroni, 2020). Funtowicz and Ravetz (1993) advocate for an ‘extended peer community’ in which laypeople and non-traditional actors participate in the evaluation of the quality of scientific knowledge, especially under conditions of uncertainty, complexity, urgency, values at stake, and when science is socially distributed and policy-relevant.

The communication of research cannot be framed reductively as a means of enhancing institutional or project visibility or demonstrating compliance with Open Science mandates; it is a constitutive element of scientific citizenship and institutional longevity. It enables RIs to articulate, synthesize, and propose their core values, establish alliances, and gain the ability to reflexively adapt to changing political and epistemic environments.

These theoretical perspectives underscore the need to reconceptualize communication within RIs as both relational and infrastructural (Fecher et al., 2021). It is through communicative actions that RIs, such as FOSSR, enact their epistemic cultures, foster productive interactions, and position themselves within broader socio-political environments. This theoretical framework guides our empirical analysis of the FOSSR project, focusing on how communication is implemented, shaped by institutional arrangements, and experienced in everyday practice.

3. Methodology

This study adopts a case study approach to investigate the evolving role of communication within an Open Science Research Infrastructure, focusing on the FOSSR project (Fostering Open Science in Social science Research) based on the implementation of an Open Cloud – a kind of meta-research infrastructure developed under Italy's National Recovery and Resilience Plan (NRRP), provides a manifold context for examining how communication can be first conceptualized and then implemented at the service of social science research. Specifically, the analysis focuses on FOSSR's dedicated communication and community-building work package, from the design phase through to the realization of strategic activities.

The case study approach is particularly suitable for exploring complex and context-dependent phenomena, such as the infrastructural role of communication. It's focused on an in-depth exploration of social processes within their contexts, allowing researchers to identify the factors driving organizational dynamics, such as situated practices, institutional logics, emerging challenges, tensions, and networks of actors.

In the FOSSR case, the communication strategy (Reale & Fabrizio, 2024) represents not only a technical function but a strategic and epistemic point of negotiation, where the meanings, roles, and boundaries are continuously constructed and negotiated within broader research and governance frameworks. The empirical material for this study was drawn from various sources, allowing a comprehensive understanding of the case. First, internal documentation – a total of 30 documents, including the project proposal (1), deliverables (5), milestones (11) and periodic reports (13) – was explored to trace the formal articulation of communication goals, strategies, and features. These documents, analysed with repeated close readings guided by some basic operational questions (e.g.: 'What was the frame for communication activities within FOSSR?' 'Which communicative

tools, practices and channels were proposed by design?' 'Which audience segments were considered, for what purpose, and which were the tailored tools to address them?' 'Which external and internal actors the specific communication practice was meant to address?'), provided insight into how communication is framed within the broader objectives of the RI and how it is embedded in institutional planning.

Second, communication plans and strategic frameworks were examined to understand the operationalization of communication within FOSSR. These materials outlined the intended audiences, key messages, channels, and evaluation metrics, offering a window into the infrastructure's communicative rationalities. Particular attention was paid to how these plans addressed stakeholder engagement and community-building, supporting the knowledge circulation principle of Open Science.

Third, the digital community – including interactions on FOSSR social media channels and newsletter subscribers – was analyzed to assess how community-building is enacted in practice. These outputs were treated as both communicative channels and performative acts that contribute to the construction of FOSSR's community identity and epistemic legitimacy.

Fourth, public events, such as workshops, webinars, and in-person meetings, were considered crucial contexts where communicative practices contribute to community building. Basic participation data were acquired: the number of events organized, the number of participants to events and their institutional affiliation category, the number of papers presented at conferences. However, these events were examined not only for their content, but also for their design and format, structure, participation dynamics, and facilitation strategies, allowing to obtain an early assessment of the extent to which FOSSR's communication practices incorporate dialogic and participatory principles.

A distinctive feature of this study lies in its collaborative and reflexive methodology, grounded on the dual positionality of the researchers as both academic analysts and actively involved practitioners, allowing for a reflexive engagement with the empirical material, facilitating the identification of tensions, contradictions, and emerging practices that might otherwise remain unnoticed. Field notes, informal conversations, and reflexive annotations were used to capture and reconstruct experiential knowledge, integrating it into the analytical process.

This collaborative approach is consistent with recent perspectives in science and technology studies (STS) and organizational sociology, calling for more engaged, situated, and participatory forms of inquiry that can appreciate the features pertaining to the performative nature of research. Research methods do not simply represent the social world, they help enact it, thus requiring reflexive and inclusive approaches in order to account for the multiplicity and complexity of social realities (Law & Urry, 2004); situated knowledge and relational practices are crucial factors to consider in order to capture all the nuances of social situations, hence methodologies that are embedded,

responsive, and attentive to the dynamics of collaboration across diverse contexts are necessary. By situating the researcher within the communicative lifeworld of the RI, the study was able to capture the lived realities of communication work, including the affective labor, institutional constraints, and creative improvisations that characterize everyday practice.

Despite its strengths, the study also shows some limitations. As a single case study, it sheds light on the in-depth dynamics of social structures; however, the findings are not intended to be generalizable in a comparative sense. Rather, they aim to provide analytical insights that can be applied to other contexts with similar characteristics. In addition, the embedded nature of the research required a careful consideration of researchers' positionality. Reflexivity was employed to mitigate these aspects, and the double role of researchers inevitably shaped the selection, interpretation, and presentation of data.

This study focuses on the internal and institutional dimensions of communication within FOSSR, while subsequent studies will be able to highlight the reception and impact of these practices among external stakeholders and the broader community.

Finally, the temporal range of the study is limited to the early phases of FOSSR's implementation, and is meant as an early indicator of design-related choices. As the infrastructure grows, its communication practices may evolve in response to new challenges, opportunities, institutional, and contextual learning. Longitudinal studies will be necessary to explore these development dynamics in order to understand the role of communication in sustaining the infrastructure over time.

4. The FOSSR Open Cloud

FOSSR's strategic aim is to integrate with a unique access point the country's nodes of key European research infrastructures: CESSDA-ERIC (Consortium of European Social Science Data Archives), SHARE-ERIC (Survey of Health, Ageing and Retirement in Europe), and RISIS (Research Infrastructure for Research and Innovation Policy Studies). Its overarching goal is to establish an Italian Open Science Cloud for the social sciences, inspired by the European Open Science Cloud (EOSC), to enhance the accessibility, interoperability, and reusability of social science data.

This kind of 'meta-infrastructure' is designed as a distributed network of data centers and virtual platforms, coordinated by the National Research Council (CNR) and involving nine institutional partners. At its completion in 2026, after 3.5 years of activity, it will provide tools and services for data collection, curation, and analysis, including support for longitudinal surveys and the possibility to explore social science issues with probability panels.

The key rationale behind FOSSR's early investment in community-building lies exactly in the recognition that research infrastructures are not only technical systems but also social constructs.

Considering RIs as sociotechnical platforms means recognising their vital role in conducting high-quality research, which often involves processing significant amounts of complex data from various sources (Watson & Floridi, 2018). They integrate technological components, human expertise, and organizational processes to enable scalable, reliable, and collaborative knowledge production (Ibidem).

The survival of any RI is contingent on the individuals operating it; however, in the context of social science RIs, which are less tangible than those in hard science, the notion of research infrastructures as socio-technical objects assumes particular significance.

The sustainability and impact of a research infrastructure depend on the robustness and vitality of the communities that use, maintain, and evolve it. Consequently, soft features such as data curation, the provision and maintenance of services to interpret and extract meanings from data, and a community controlling and validating the quality of data and the production of high-quality research are at the heart of Social Science RIs.

Accordingly, FOSSR has prioritized the development of communicative and participatory mechanisms that support building of a network of knowledge exchange and infrastructure 'co-building' by means of collaborative practices among its stakeholders. This approach acknowledges that Open Science is not merely a matter of open data or open access, but a cultural and institutional transformation that requires sustained engagement and mutual learning.

FOSSR's target users span a broad ecosystem: academic researchers, public administrations, policymakers, civil society organizations, and even engaged citizens. This diverse stakeholder landscape reflects the project's commitment, alongside to provide robust datasets and services to the scholarly community, to democratize the access to scientific knowledge and to foster a widening of the participation in research processes. By lowering technical and institutional barriers to data access and analysis, FOSSR seeks to empower a wider range of actors to engage with and contribute to social science research, in addition to exploit evidence-based outputs.

5. Communication Strategy: From Visibility to Engagement

FOSSR's communication strategy is based on a model comprising four levels of engagement that are dynamically related: Inform, Involve, Collaborate, and Empower. This multilevel model (Yang & Shen, 2015) is implemented in the RISIS communication approach, which has successfully structured its user engagement around similarly graduated levels of interaction. Different audiences are progressively activated through

training, seminars, and policy dialogues (Fabrizio et al., 2023). The concept of this model reflects a progressive deepening of interplay between the infrastructure and its stakeholders, allowing for the overcoming of conventional dissemination towards co-creation and shared governance.

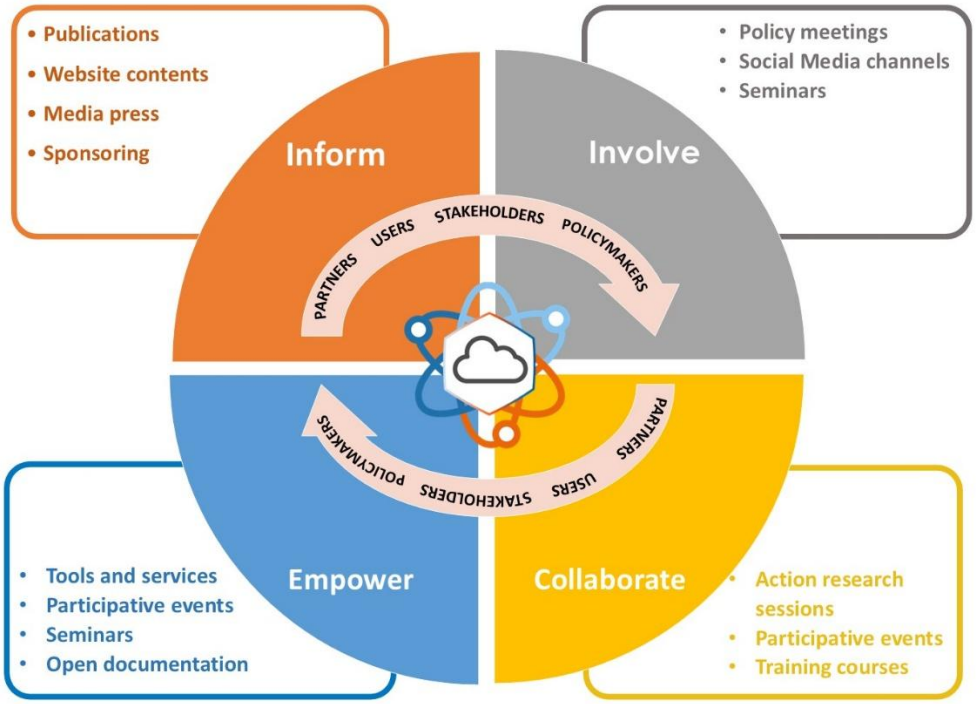


Figure 1: FOSSR Engagement Chart (FOSSR Communicaton Plan, January 2023)

The social actors that comprise the FOSSR infrastructure are understood as comprising different actors, who are involved according to their roles and levels of engagement. The *actual users* are defined as individuals or organisations that are already utilising FOSSR resources, such as scholars in the social sciences, data scientists, and doctoral holders. The *potential users* are groups that may benefit from FOSSR but are not yet fully engaged, including public managers, NGOs, and innovation companies. Finally, *producers* are defined as actors involved in building or maintaining the infrastructure, such as the FOSSR partners, other RIs, and statistical officers.

This classification echoes the tripartite division adopted in RISIS communication strategy (scholars, stakeholders, policymakers), allowing FOSSR to benefit from a tested segmentation that guided targeted communication actions (Ibidem).

The actors are further grouped into four main categories. The target user base comprises academics and data experts who actively use or have the potential to use the infrastructure. Stakeholders include organisations and institutions with an interest in the infrastructure's outputs (e.g., open data repositories, foundations). With the term 'policymakers' within FOSSR, we refer to public agencies, EU officers, and political actors

who may shape or benefit from research-informed policy; 'FOSSR Partners', finally, are core institutions that develop and operate the infrastructure.

A significant proportion of these actors are affiliated with multiple categories, thereby reflecting the interconnected, multi-actor nature of the research infrastructure community.

The success of the engagement strategy depends on several organizational conditions. Internally, FOSSR has established a cross-institutional communication team with expertise in science communication, digital media, institutional communication and stakeholder engagement. Regular coordination meetings, shared editorial calendars, and collaborative tools support the integration of communication across work packages. Training and capacity-building for communication staff are also prioritized, recognizing that effective engagement requires both technical skills and reflexive awareness.

These internal practices reflect and expand the RISIS experience of a centralized but collaborative communication model, which integrates various expertise and ensures a coherent project voice across channels.

FOSSR’s strategy also addresses the challenges of epistemic diversity and digital inequalities. As Grand et al. (2015) argue, science communication must account for the plurality of knowledge systems and the uneven distribution of digital access and skills, challenges that FOSSR addresses by adopting inclusive language, offering multilingual content, and carefully selecting communication channels to facilitate participation.

The inclusive and accessible communication principles align with the lessons learned in RISIS, especially during the post-pandemic phase, when online accessibility and multilingual dissemination were crucial in reaching broader, more diverse audiences.

6. Outcomes and Reflections

At approximately the midpoint of the project duration, it is possible to present and analyse some mid-term results, deriving from the indicators collected during and after communicative actions (both in person events and digital activities). These outcomes are useful in terms of providing some early indications regarding the community-building-related objectives and the ability to foster participation and support Open Science.

From a quantitative point of view, FOSSR’s outreach and community-building efforts have shown increasing engagement on a range of diverse channels selected in the communication strategy.

With regard to digital communication, FOSSR’s visibility on social media appears well established, even though the communication style is shaped on very institutional features, publishing information and details mainly about events and outcomes of the project, hence stimulating few bi-directional exchanges (Giuffredi et al., 2024). These platforms are important as informal communication channels to inform the community

about the day-by day life of the project, particularly for the research community and the digitally active stakeholders. The project's YouTube videos have garnered nearly 1,500 views, while the Facebook and LinkedIn channels are experiencing consistent growth in their followers. Given necessary adjustments for the different longevity and nationality/internationality of the platforms, the figures align with the order of magnitude observed in analogous social science research infrastructures at the European level.

The infrastructure’s mailing list comprises 826 subscribers at the time of writing, primarily acquired in the context of organised events, suggesting a solid base of recurring interest. The events attracted over 1,000 attendees, reflecting that the communication strategy has succeeded in creating a shared interest in the project's activities. With regards to the dissemination of scientific outputs, over 4500 downloads from Zenodo suggest that FOSSR’s digital products are being actively accessed and downloaded by the relevant target audience, a key indicator of knowledge circulation and impact.

We integrated quantitative evidence from the participation to events organised by the project (summarised in Table 1) to substantiate the interpretation of community-building processes. Across training and communication events, we registered 1213 participations from 418 different organisations, demonstrating that engagement extended far beyond the project consortium. A stable core group of participants attended between 10 and 12 events each, signalling the emergence of an active community rather than one-off information transfer. Participation was balanced by gender (604 F, 578 M) and spanned multiple profiles (researchers, PhD students, professors), with events distributed across training, networking and dissemination, showing multidirectional interaction patterns. The annual growth – from 67 participants in 2022 to 746 in 2024 – further confirms the consolidation of this community. The challenges identified qualitatively are also reflected quantitatively: for example, incomplete preference-of-participation data (935 missing entries) and the wide organisational dispersion (418 institutions) illustrate structural difficulties in standardising engagement pathways.

After normalising all university-related entries, universities represent the second largest group of participants (276 entries), immediately after the CNR. Their internal heterogeneity is remarkable: participation comes from more than forty Italian and international universities, with no single institution dominating. The breakdown shows a long tail pattern, with medium and small universities participating alongside larger ones (e.g., Catania, Bologna, Roma Tre, Molise, Trento, Urbino, Sapienza). This confirms that community building extended across a highly decentralised academic ecosystem and did not concentrate engagement in only a few institutions.

The analysis of the available data shows that 525 unique participants took part in the project, revealing a strongly uneven distribution of attendance that reflects different roles and levels of engagement within the emerging community. The distribution is highly asymmetric: while a large share of participants attends only one or two events, a smaller

but very active core engages repeatedly throughout the project. Several individuals appear in more than ten events, with peaks of 14 participations. This core group represents the most dynamic segment of the community, surrounded by an intermediate layer of participants who join between five and ten events and constitute a stable backbone that follows the project over time. Alongside these groups, the majority of participants attend only a limited number of events, contributing to broadening the community and diversifying the stakeholder landscape, even if they are less involved in co-design and collaborative activities.

A closer look at the types of events attended further clarifies the nature of this distribution. Participation is relatively balanced across the three main formats – Training (301 participations), Networking (359), Policy dialogue (168), Service Demo (108) and Dissemination (277) – indicating that stakeholders are not only receiving information but also engaging with one another and building competencies. Activities such as Policy Dialogues and Service Demonstrations attract fewer participants, yet they tend to involve the most committed members of the community, who return across multiple events and contribute to shaping discussions and identifying needs. Taken together, these patterns point to a multi-layered engagement structure and show that the project has been able to mobilise a broad audience while simultaneously consolidating a core group of highly active participants who drive collaborative and co-creative processes.

Indicator	**Value/Typology**
Total participations	1213
Unique participants	525
Organisations involved	418
Gender distribution	604 F – 578 M (25 n.a.)
Events participants[1]	Training (301) Networking (359) Dissemination (277) Policy dialogue (168) Service Demo (108)
Annual growth of the FOSSR community	67 (2022) → 746 (2024)

Table 1: Indicators of participation in events organised by the project. Data available at 31/12/2024.

This level of involvement points to a sustained demand for knowledge and capacity-building in Open Science practices, particularly in the field of social sciences. To further support the reinforcing of the community around FOSSR and the infrastructure's relational ethos, these sessions were designed not only as moments of knowledge transfer but also as spaces for dialogue and mutual learning.

[1] Participants profiles: Researchers, PhD students, Professors.

Policy engagement was designed within FOSSR as a specific area of emphasis, identified as a primary target social sector for the research infrastructure to achieve a substantial societal impact, sharing the Open Science approach among relevant communities. FOSSR has hosted dedicated policy sessions with 169 overall registered participants, and related materials – policy briefs and policy-makers sessions presentations – have been downloaded over 500 times. These figures suggest that the FOSSR work at the science-policy interface is starting to stimulate interest and attract an interested audience.

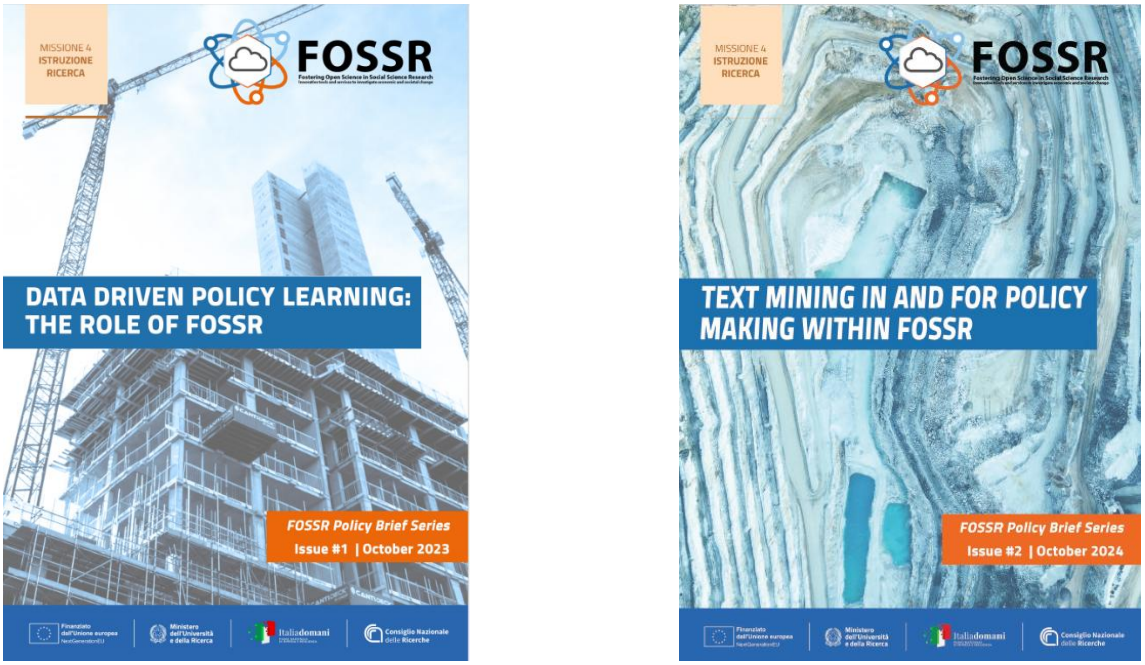


Figure 2. Policy Brief Issues

These indicators suggest the emergence and early stabilisation of a community built around FOSSR's datasets, services and research advancements and outputs, with the peculiar physiognomy of an interconnected community. The infrastructure, indeed, appears not to be only reaching diverse audiences but also facilitating sustained interaction among them in terms of scientific, technical and strategic reflexive exchanges on the themes and developments of the project, laying the grounds for long-term support and infrastructure co-creation.

The qualitative analysis was particularly focused on the design features of communication, with reference to its positioning in relation to community-building and to the support of knowledge circulation, which is notably relevant in Open Science contexts. The communication plan, developed in consideration of prior experiences with European-wide infrastructures (RISIS in particular), included a full acknowledgement of the crucial relevance of existing communities, primarily composed of scholars and stakeholders, and of the communicative practices targeted at supporting engagement with the new RI. In

addition, the importance of internal and external networking, as well as the active involvement of scholars in the field of social sciences, was emphasised as a key approach to be pursued. A key target group in terms of societal impact, policy-makers, were targeted with specific activities and communication products (policy-makers sessions and a policy brief series), with the attention to employ a suitable language – rigorous but not cryptic –, pointing out pragmatic and reusable outputs.

A particular attention was devoted to the realisation of a coordinated and recognisable image for the RI, as well as for transparent, coherent and continuous communication practices, responsiveness to feedback on all channels, and the consistent attention on communication coherence, both in terms of style and contents, in all activities with the public. Another key outcome related to the development of a recognised network among all the diversified actors of the community refers to the establishment of trust in the audience, supported by field observations on the establishment of recurring interactions (in terms, for example, of returning participants to trainings, informal feedbacks at public events). A trustworthy community underpins users' willingness to contribute data, participate in proposed events, and provide feedback to governance choices, and possibly contribute to sharing the infrastructure within their own networks.

However, the implementation of the communication strategy has not been without challenges, which can in part be understood as intrinsic to the innovative and transformative character of Open Data research infrastructures in the social sciences, requiring an effort of understanding and adaptation to all members of the community. A persistent challenge, testified by field observations, concerns the time investment required for meaningful engagement, in particular for what concerns participatory formats that demand significant preparation, facilitation, and subsequent follow-up from the communication group, in addition to a willingness from the part of participants to actively contribute and reflexively adapt. This is particularly acute in the context of academia, where the significance of communication is still persistently underestimated concerning technical or scientific work.

Institutional expectations also pose challenges, particularly regarding the exploratory research work of the communication staff. While funders and host institutions increasingly recognize the importance of communication, they may still prioritize visibility metrics and traditional one-way communication over deeper, and potentially game-changing, forms of engagement. This can create pressure to produce high volumes of content at the expense of a real engagement of the audience, oriented to long-term community-building. Moreover, disciplinary asymmetries persist; while some fields within the social sciences are well-versed in participatory methods, others may be less familiar or less inclined to engage in co-creative practices. In addition, any inter- or trans-disciplinary confrontation requires additional effort of the facilitation staff, since it's necessary to consider and engage with implicit disciplinary assumptions regarding

methodologies and the research process and diverse backgrounds influencing the meanings assigned to research.

Despite these tension knots, it's possible to witness encouraging signs of community consolidation. Field observation corroborates mid-term outcomes on the fact that informal networks have begun to form among participants in FOSSR events, and several collaborative initiatives have emerged from these interactions. These include joint training proposals, shared data projects, and cross-institutional working groups. Such developments suggest that the FOSSR approach is not only disseminating knowledge but also catalysing new forms of collective action.

These emerging dynamics echo those observed during the RISIS project, where sustained investments in integrated communication and community engagement laid the groundwork for a lasting, interdisciplinary network. FOSSR, by drawing upon this legacy and adapting it to the context of Open Science in the social sciences, is demonstrating the potential of research infrastructures not only to produce and share knowledge but to convene, empower, and transform communities of practice.

7. Discussion

The analysis of the design and mid-term results of the FOSSR communication strategy underscores that the understanding of communication within Open Science research infrastructures should add to a conventional institutional communication perspective, aimed at informing the public, boosting the visibility of the project and reinforcing the strategic and promotional objectives of the institution, the awareness of its infrastructural function related to community-building necessary to long-term sustainability, overcoming the persistent representation as a mere support activity to scientific activities. This reconceptualization has significant implications for how RIs are designed, governed, and evaluated.

Considering communication infrastructural is to recognize its role in enabling the establishment and maintenance of lively, continuous, and coherent relations among all the actors involved within the project, supporting coordinated research activities, full stakeholders' engagement and societal robustness. Communication practices – whether in the form of documentation, dialogue, or digital mediation – serve as the fundamental binding elements uniting the diverse components of an RI, considering both internal and external actors. They facilitate interoperability and collaboration not only at the technical level but also at the epistemic and organizational levels. In FOSSR, communication has served to align expectations, translate between disciplinary languages, and mediate with institutional logics.

This infrastructural role also entails a shift in how communication is understood, valued, and resourced. Rather than being treated as a service function or a PR office, as it often

happens especially with institutional approaches (Claessens, 2018), communication should be integrated into the strategic architecture of RIs. This includes dedicated staffing, strategic planning in close connection with the scientific and organisational coordination functions, and mechanisms for continuous learning and adaptation. FOSSR's experience shows that such integration can enable communication to shift from a service role to a co-creative practice, where stakeholders and the research community are not passive recipients but active contributors to the infrastructure's development, thereby laying the groundwork for long-term relevance.

Such a change in perspective has wide implications for the identities and boundaries of research. As communication becomes more participatory and dialogic, and in parallel as research becomes hybrid, contextual and commissioned (Benessia et al., 2012; Gibbons et al., 1994; Ziman, 2000), traditional distinctions between producers and users of knowledge are challenged. Researchers are called upon to engage in meaningful conversations with publics, policymakers, stakeholders and practitioners not only as audiences but as collaborators, while the public is often asked to participate in co-designing research objectives and governance, especially on societally relevant topics (Bucchi & Trench, 2021). This reconfiguration of roles is an ongoing process, driven by the increase in importance of social contexts of knowledge production with features of interdisciplinary and intersectoral hybridisation, as it happens in Research Infrastructures or transdisciplinary research endeavours, which in themselves require notable efforts of communication (Davies & Horst, 2016; Harris et al., 2024). Perceiving the blurring of roles can be unsettling, particularly in disciplines where the authority of expertise is closely guarded. Yet it also opens up new possibilities for relevance, legitimacy, and impact (Cash et al., 2002; Marenko, 2021).

The transformation of communication practices reflects and contributes to the transformation of research itself, rethinking the processes, relationships, and values that underlie knowledge production. Communication is central to this transformation because it intrinsically has a relational personality, positioning itself in a suitable position to mediate and contribute to shaping the very conditions under which knowledge is produced, shared, and institutionalized (Davies & Horst, 2016).

The FOSSR community-oriented perspective, which is described here in its early stages of implementation and is considered an ideal final aim, is able to amplify and support the development of features that are already visible within the experience in the current RI.

The attention to reference communities, although foundational to science communication (especially for what concerns the inclusion of different epistemic communities or lay knowledge producers, Collins & Evans, 2002; Wynne, 1992), is recently losing ground to strategic orientation for communicative practices, especially within research institutions (Kessler et al., 2022; Nisbet & Markowitz, 2016; Orthia et al., 2021). Our results show that in the case of a Social Science Research Infrastructure, the actual community,

grounding the RI's same existence and survival, extends to all actors involved in the RI, abolishing the traditional clear demarcation between science and society and promoting all actors to co-responsibility in their diverse roles in the infrastructure. The objective of promoting all 'stakeholders' (with all the limitations beared by this term, as argued in Reed et al., 2024) to active actors and possibly co-builders of the RI, overcoming the conventional distinction between the inside and the outside of the research endeavour, was addressed by design in FOSSR in a variety of ways, and especially within the diverse committees – stakeholders, strategic, scientific, etc. – involved in the governance of the project. The contributions of experts and community members, who are often excluded from direct participation in defining and refining the project, are considered essential for the collaborative development of the project's evolution and outcomes in FOSSR.

Such a transformative effort necessitates a community effort to understand and act on a reflective paradigm shift. Drawing from the reflections of scholars on the changes in the contemporary system of knowledge production, both from STI and from STS, FOSSR proposed some activities specifically targeted to foster in the involved members of the community a rethinking of the implicit consolidated attitudes towards the research system and its positioning in society, also to positively integrate any tension line during the project development which could be understood as lying on different pre-existent views of research clashing with others'.

FOSSR's experience offers several lessons for other Open Science infrastructures, as well as development paths to be incorporated in the second phase of FOSSR and in the post-project evolution.

First, communication strategies must acknowledge the delicate and ever-changing interface between research and society on which they lie, be context-sensitive, and be attuned to the specific epistemic cultures, stakeholder landscapes, and institutional constraints of each RI. The communication group is of pivotal importance in identifying gaps or deficiencies in frame-setting, information, process-related or organisational exchanges, and in facilitating the emergence and confrontation of these issues.

Second, communication strategies should include by design spaces of reflexivity, opening to the emergence and mediation of tacit expectations and concerns, and allowing informal exchange among the diverse actors participating in the infrastructure. Furthermore, the projectual strategy must be sufficiently flexible to incorporate the outcomes of these reflections, allowing for learning from feedback and adaptation to change.

Finally, the case of FOSSR suggests that communication, as a strategic infrastructural function that grounds and supports the building of communities, can serve as a lever for institutional change in the research sector. By embedding participatory practices into the infrastructure's design, FOSSR has begun to shift organizational cultures and expectations within the existing community. Although this is, by its own nature, a slow

and ever-evolving process, it holds promise for building more resilient, responsive, and democratic research systems.

8. Conclusion

Drawing on sociological and STS perspectives, we have argued that communication in RIs is well positioned to evolve its understanding towards the valorization of relational and infrastructural functions, enabling collaboration, sustaining the mediation of different epistemic cultures, and supporting the transformation of research institutions.

FOSSR's communication and engagement strategy, structured around the levels of informing, involving, collaborating, and empowering, is designed to allow communication to move beyond visibility to foster meaningful awareness and long-term support. Through a combination of digital tools, participatory formats, and reflexive practices, FOSSR is in its early stages of building a diverse extended research community around the infrastructure. Early outcomes indicate the stabilization of a network and the emergence of collaborations, even as challenges related to navigating a transformative process, both for researchers and the community, persist.

The development of this new perspective has broader relevance for the governance and design of Open Science research infrastructures, especially for those sharing with FOSSR the focus in the field of social sciences. A key suggestion is that communication should not be viewed as an auxiliary function, but rather as a core strategic component of the infrastructure. This requires investment in knowledge and skills oriented to inter- and trans-disciplinary exchanges, enhanced coordination, and the practice of inclusive and participative actions, as well as a willingness to rethink traditional roles and boundaries in research.

Looking ahead, future work should focus on evaluating the long-term impacts of communication strategies on community-building, knowledge co-production, and institutional sustainability, especially in comparison with analogous Research Infrastructures in different contexts. Comparative studies across RIs could help identify common issues and best practices, share reflections and research on the processes, and develop cross-infrastructure models for participatory communication oriented to community-building. As Open Science continues to evolve, the ability of infrastructures to serve as a lever of change, communicating effectively and inclusively, will be central to their long-term societal impact, relevance, and sustainability.

Acknowledgement

This work was supported by FOSSR (Fostering Open Science in Social Science Research), funded by the European Union – NextGenerationEU under NRRP Grant agreement n. MUR IR0000008. The content of this article reflects only the author's view. The European Commission and MUR are not responsible for any use that may be made of the information it contains.

References

Benessia, A., Funtowicz, S., Bradshaw, G., Ferri, F., Ráez-Luna, E. F., & Medina, C. P. (2012). Hybridizing sustainability: towards a new praxis for the present human predicament. *Sustainability Science*, *7*(S1), 75–89. https://doi.org/10.1007/s11625-011-0150-4

Benessia, A., Funtowicz, S., Guimarães Pereira, Â., Ravetz, J. R., Saltelli, A., Strand, R., & van der Sluijs, J. P. (2016). *The Rightful Place of Science: Science on the Verge*. Consortium for Science, Policy and Outcomes at Arizona State University.

Bodmer, W. F. (1985). *The Public Understanding of Science: Report of a Royal Society Ad Hoc Group Endorsed by the Council of the Royal Society.*

Bucchi, M., & Neresini, F. (2008). Science and public participation. In E. J. Hackett, O. Amsterdamska, M. Lynch, & J. Wajcman (Eds.), *The handbook of science and technology studies* (3rd ed.). MIT Press.

Bucchi, M., & Trench, B. (2021). Rethinking science communication as the social conversation around science. *Journal of Science Communication*, *20*(03), Y01. https://doi.org/10.22323/2.20030401

Cash, D., Clark, W. C., Alcock, F., Dickson, N. M., Eckley, N., & Jäger, J. (2002). *Salience, Credibility, Legitimacy and Boundaries: Linking Research, Assessment and Decision Making* (KSG Faculty Research Working Paper Series). http://ssrn.com/abstract_id=372280

Cerroni, A. (2006). *Scienza e società della conoscenza*. UTET Università.

Cerroni, A. (2020). *Understading the knowledge society: a new paradigm in the sociology of knowledge.* Edward Elgar Publishing.

Claessens, M. (2018). Research institutions: neither doing science communication nor promoting ‘public’ relations. *Journal of Science Communication*, *13*(03), C03. https://doi.org/10.22323/2.13030303

Collins, H. M., & Evans, R. (2002). The Third Wave of Science Studies. *Social Studies of Science*, *32*(2), 235–296. https://doi.org/10.1177/0306312702032002003

Cortassa, C. (2016). In science communication, why does the idea of a public deficit always return? The eternal recurrence of the public deficit. *Public Understanding of Science*, *25*(4), 447–459. https://doi.org/10.1177/0963662516629745

Davies, S. R., & Horst, M. (2016). Science Communication - Culture, Identity and Citizenship. In *Science Communication* (Vol. 31, Issue 1). Palgrave Macmillan. https://doi.org/https://doi.org/10.1057/978-1-137-50366-4_5

Fabrizio, S., Reale, E., & Spinello, A. O. (2023). *Policy Brief, Issue 15/ Building a stronger community in STI studies for effective impact: the experience of RISIS (Versione 1). Zenodo*. https://doi.org/Zenodo.https://doi.org/10.5281/zenodo.11276785

Fecher, B., Kahn, R., Sokolovska, N., Völker, T., & Nebe, P. (2021). Making a Research Infrastructure: Conditions and Strategies to Transform a Service into an Infrastructure. *Science and Public Policy*, *48*(4), 499–507. https://doi.org/10.1093/scipol/scab026

Funtowicz, S., & Ravetz, J. (1993). Science for the post-normal age. *Futures*, *25*(7), 739–755. https://doi.org/10.1016/0016-3287(93)90022-L

Gibbons, M., Limoges, C., Nowotny, H., Schwartzman, S., Scott, P., & Trow, M. (1994). *The New Production of Knowledge: The Dynamics of Science and Research in Contemporary Societies*. SAGE.

Giuffredi, R., Grasso, V., & L'Astorina, A. (2024). Web-based science communication at Research Institute level: balancing dissemination, dialogue and promotion in a major Italian scientific institution. *Frontiers in Communication*, *9*. https://doi.org/10.3389/fcomm.2024.1427033

Grand, A., Davies, G., Holliman, R., & Adams, A. (2015). Mapping Public Engagement with Research in a UK University. *PLOS ONE*, *10*(4), e0121874. https://doi.org/10.1371/journal.pone.0121874

Harris, F., Lyon, F., Sioen, G. B., & Ebi, K. L. (2024). Working with the tensions of transdisciplinary research: A review and agenda for the future of knowledge co-production in the Anthropocene. *Global Sustainability*, *7*. https://doi.org/10.1017/SUS.2024.11

Hecker, S., Haklay, M., Bowser, A., Makuch, Z., Vogel, J., & Bonn, A. (2018). *Citizen Science: Innovation in Open Science, Society and Policy*. UCL Press. https://doi.org/10.14324/111.9781787352339

Irwin, A. (2015). Citizen science and scientific citizenship: same words, different meanings? In B. Schiele, J. L. Marec, & P. Baranger (Eds.), *Science Communication Today29* (pp. 29–38). Presses Universitaires de Nancy.

Jasanoff, S. (2003). Technologies of humility: citizen participation in governing science. *Minerva*, *41*(3), 223–244. https://doi.org/10.1023/A:1025557512320

Jasanoff, S. (2004a). Science and citizenship: a new synergy. *Science and Public Policy*, *31*(2), 90–94. https://doi.org/10.3152/147154304781780064

Jasanoff, S. (2004b). *States of Knowledge: The Co-Production of Science and the Social Order*. Routledge.

Jasanoff, S. (2005). *Designs on Nature: Science and Democracy in Europe and the United States.* Princeton University Press.

Kessler, S. H., Schäfer, M. S., Johann, D., & Rauhut, H. (2022). Mapping mental models of science communication: How academics in Germany, Austria and Switzerland understand and practice science communication. *Public Understanding of Science*, *0*, 096366252110657. https://doi.org/10.1177/09636625211065743

Knorr-Cetina, K. (1999). *Epistemic Cultures: How the Sciences Make Knowledge*. Harvard University Press.

Law, J., & Urry, J. (2004). Enacting the social. *Economy and Society*, *33*(3), 390–410. https://doi.org/10.1080/0308514042000225716

Marenko, B. (2021). Stacking Complexities: Reframing Uncertainty through Hybrid Literacies. *Design and Culture*, *13*(2), 165–184. https://doi.org/10.1080/17547075.2021.1916856

Miller, J. D. (2004). Public Understanding of, and Attitudes toward, Scientific Research: What We Know and What We Need to Know. *Public Understanding of Science*, *13*(3), 273–294. https://doi.org/10.1177/0963662504044908

Molas-Gallart, J., & Tang, P. (2011). Tracing 'productive interactions' to identify social impacts: an example from the social sciences. *Research Evaluation*, *20*(3), 219–226. https://doi.org/10.3152/095820211X12941371876706

Nisbet, M. C., & Markowitz, E. (2016). *Strategic science communication on environmental issues - Commissioned White Paper in Support of the Alan Leshner Leadership Institute American Association for the Advancement of Science*.

Nisbet, M. C., & Scheufele, D. A. (2009). What's next for science communication? promising directions and lingering distractions. *American Journal of Botany*, *96*(10), 1767–1778. https://doi.org/10.3732/ajb.0900041

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers, C. D., Chin, G., Christensen, G., Contestabile, M., Dafoe, A., Eich, E., Freese, J., Glennerster, R., Goroff, D., Green, D. P., Hesse, B., Humphreys, M., … Yarkoni, T. (2015). Promoting an open research culture. *Science*, *348*(6242), 1422–1425. https://doi.org/10.1126/SCIENCE.AAB2374,

Nowotny, H., Scott, P., & Gibbons, M. (2001). *Re-thinking science: Knowledge and the public in an age of uncertainty*. Polity Press.

Orthia, L. A., McKinnon, M., Viana, J. N., & Walker, G. (2021). Reorienting science communication towards communities. *Journal of Science Communication*, *20*(03), A12. https://doi.org/10.22323/2.20030212

Reale, E., & Fabrizio, S. (2024). *FOSSR Deliverable 9.1 - Dissemination and communication plan (Version 1). Zenodo*.

Reed, M. S., Merkle, B. G., Cook, E. J., Hafferty, C., Hejnowicz, A. P., Holliman, R., Marder, I. D., Pool, U., Raymond, C. M., Wallen, K. E., Whyte, D., Ballesteros, M., Bhanbhro, S., Borota, S., Brennan, M. L., Carmen, E., Conway, E. A., Everett, R., Armstrong-Gibbs, F., … Stroobant, M. (2024). Reimagining the language of engagement in a post-stakeholder world. *Sustainability Science*, *19*(4), 1481–1490. https://doi.org/10.1007/s11625-024-01496-4

Star, S. L., & Ruhleder, K. (1996). Steps Toward an Ecology of Infrastructure: Design and Access for Large Information Spaces. *Information Systems Research*, *7*(1), 111–134. https://doi.org/10.1287/ISRE.7.1.111

Trench, B. (2008). Towards an Analytical Framework of Science Communication Models. In *Communicating Science in Social Contexts* (pp. 119–135). Springer Netherlands. https://doi.org/10.1007/978-1-4020-8598-7_7

UNESCO. (2021). *UNESCO Recommendation on Open Science*. https://doi.org/10.54677/MNMH8546

Watson, D., & Floridi, L. (2018). Crowdsourced science: sociotechnical epistemology in the e-research paradigm. *Synthese*, *195*(2), 741–764. https://doi.org/10.1007/s11229-016-1238-2

Wynne, B. (1992). Misunderstood misunderstanding: social identities and public uptake of science. *Public Understanding of Science*, *1*(3), 281–304. https://doi.org/10.1088/0963-6625/1/3/004

Wynne, B., Felt, U., Callon, M., Gonçalves, M., Jasanoff, S., Jepsen, M., Joly, P.-B., Konopasek, Z., May, S., Neubauer, C., Rip, A., Siune, K., Stirling, A., & Tallacchini, M. (2007). *Taking European Knowledge Society Seriously. Report of the Expert Group on Science and Governance to the Science, Economy and Society Directorate, Directorate-General for Research, European Commission.*

Yang, R. J., & Shen, G. Q. P. (2015). Framework for Stakeholder Management in Construction Projects. *Journal of Management in Engineering*, *31*(4). https://doi.org/10.1061/(ASCE)ME.1943-5479.0000285

Ziman, J. (2000). *Real Science: What it Is and What it Means*. Cambridge University Press.

Health Data Circulation in France: Between Public Interest and Privacy Enhancing Technologies

Margo Bernelin

CNRS, France

DOI 10.3217/978-3-99161-062-5-016, CC BY 4.0

Abstract. Within the healthcare system, the promises of Privacy-enhancing technologies (PETs) have attracted considerable attention to the point that, in France, personal health data cannot be used for anything other than care if it is not protected by such Technologies. This movement toward exploring more closely 'data circulation-privacy-friendly' solutions emerged about ten years ago in a context where the State was willing to encourage health data circulation for medical research. Indeed, in France, the most important health databases are operated by the State and have the advantage of being comprehensive in terms of population. In a bidding war with other States that were also willing to open their databases for research, the French Government introduced a bill to make the national databases accessible for research. To obtain support for the bill, some Members of Parliament and Senators, but also the French Health Ministry kept putting forward the benefits of Privacy Enhancing-Technologies in protecting health data and focussed on the public interest dimension of sharing health data for research purposes. Analysing the legal landscape and discourse, this paper demonstrates that since 2016, Privacy Enhancing-Technologies have been a key factor in authorising data access alongside the public interest to conduct research. Rather than closing any debate on data privacy, it has actually opened new questions on the efficacy of Privacy Enhancing-Technologies and on what their scope should include.

1 Introduction

With the discussion and recent publication of the Health Data Space Regulation by the European Union (march 2025)[2], the accessibility of health data for research has gained increased attention (Shabani 2022, Aufrechter 2025). Indeed, the new regulation seeks to promote the digitalisation of medical records and other health data by ensuring that all EU citizens have access to an electronic version of these records, or at least part of them (Marelli 2023, Horgan 2022, Mergelin 2024.). Moreover, the regulation requires health data holders, such as hospitals or pharmaceutical companies, to allow others to access their data for research purposes (de Grove Valdeyron 2024). Such a requirement has created significant hope for data sharing for public good but also concerns regarding privacy (Bernelin 2024).

Health data are defined by the 2016 General Data Protection Regulation (GDPR) as 'personal data related to the physical or mental health of a natural person, including the provision of health care services, which reveal information about his or her health status'[3]. Such data, including medical files, medical screening reports, health insurance data, prescription sheets or event data from health-related IoT devices, are sensitive by nature. Indeed, they reveal very intimate information about one's health issues that he or she might not want to share with anyone. Moreover, the data does not only relates to one's health but also to their family members, as medical history and some hereditary health components or diseases are also included in medical files. In addition, medical records often also include information about employment, family life, hobbies, housing situation or any other element that might have an impact on health.

The variety of data as well as what they can reveal (diseases or their risk of happening, injuries, disabilities, sexual orientation or experiences of abuses) require their strong protection against undue access. In this regard, it is needless to say such data could be used as discriminatory weapons with regards to access to employment or to services such as loans (for instance see Garcia, 2024 p. 2062). Such data can also offer crucial information for malicious individuals that will use it for illegal activities such as phishing, catfishing or even blackmailing (Kleinman 2020, or CybersecurityAsia.net 2024). Their protection is therefore crucial. One way to achieve it is to keep them secret. Physicians, or General Practitioners depending on your location, are already subject to medical secrecy/confidentiality obligations. In France, not only is it a deontological obligation but also a legal one, which requires health care professionals and any other professional

[2] Regulation (EU) 2025/327 of 11 February 2025 on the European Health Data Space and amending Directive 2011/24/EU and Regulation (EU) 2024/2847.

[3] art. 4 (15) Regulation 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).

accessing health information (such as data analysis in hospital or administrative agents) to keep them secret or to face criminal sanctions[4]. To lawfully share patient data with someone else, medical professionals must have compelling reasons such as, in case of an emergency, to save one's life or with the patient's consent. In France, the judges have recently ruled that sharing identifiable health data for teaching purposes with medical students is not legitimate as such and requires, to be lawful, the patient's consent. In this instance, judges held the medical practitioner liable for this breach of confidentiality[5].

Sharing health data is, nevertheless, considered crucial for various reasons. It is, indeed, important to share patient data among a medical team in order to ensure the provision of optimal care for a patient. It is also important, for public health purposes, to have data about a population or specific population groups in order to design and implement health plans and, for public authorities, to make informed decisions. It is also paramount to exploit health data for research purposes, whether the data was collected directly for research or for another purpose, such as patient care. In this regard, various and large datasets are useful, for instance, to predict a patient's trajectory by comparing it with that of another patient. Health data can also be crucial to observe drugs side effects, to train Artificial Intelligence-based Medical Devices used in patient care and more generally for applying *big data* approaches in health to foster the formulation of new hypothesis, detect diseases before they appear or to provide for more personalised treatment (Bernelin and Desmoulin 2020, Mercier 2020). However, processing and sharing health data inherently involve privacy risks.

In this context, ten years ago, the French Parliament authorised the creation of a large and centralised (in Paris) health 'super' database (Information System - IS) in order to make health data available for research purposes. By adapting medical confidentiality rules and data protections laws, the Acts attracted much attention (Bossi Malafosse 2016 and 2016a; Debiès 2016, Desmoulin-canselier 2018, Devillier 2017). Indeed, they have been widely analysed by the legal literature, which has extensively studied the type of data being made available for research (Bernelin 2020, p.26), patients' individual rights (Debiès 2016), the public interest in the circulation of health data (Péchillon 2015, Teller 2022, Pailhès 2018) and the governance mechanism of this Information System (Supiot, 2020). The question of how to protect privacy in this setting was largely left on the back burner, legal scholar envisioning this dimension more as a technical issue that, indeed, required mention (Devillier 2017) but did require deeper analysis and questioning. However, since then and with the entry in force of the General Data Protection Regulation (GDPR), privacy issues have emerged as a focal point in the case law concerning health

[4] Article L1110-4 French Public health Code and article 226-13 French Criminal Code.

[5] Strasbourg Tribunal Administratif, 5th Cahmber, July, 4th 2024 (n°2207563).

data sharing for research[6] with the literature questioning the GDPR application to health-related research on data (Marelli *et al*. 2021). Such a situation begs the following questions: What was the role played, at the time, by Privacy Enhancing-Technologies (PETs) for the regulatory creation of a centralised health Data Space in France? How has this role materialised in law? How, post-GDPR and the creation of multiple Health Data Spaces, are these Privacy Enhancing-Technologies referred to by Judges when other socio-ethical issues, such as public interest in research, are in balance with health data Access?

Answering these questions is crucial to understand how acts were enacted, their wording and their current application. Such answers provide material to support a critical analysis of the regulation of health data access, which appears to be caught between technological imaginaries of protection and public interest in research. To answer those questions, one has to return to the very beginning in order to dissect how the Acts were influenced by privacy issues. Indeed, the legal discourse in the parliamentary debates provides privileged insight into how technological objects and the risk associated with them were framed and balanced against socio-ethical values. Their study is paramount for a better understanding of the rules governing health data access prior to the creation of many Health Data Spaces (Hoeyer *et al.* 2024), including the European Health Data Space (EEDS, March 2025), and for analysing their consequences today for health data circulation. Such analysis is missing from the scholarship that has rather primarily focused on the recent EEDS and its complex articulation with the GDPR and the European Union Artificial Intelligence Act 2024 (De Grove-Valdeyron ed. 2025, Quinn *et al.* 2024), on the utilitarian inspiration behind its regulatory model (Lianos 2025), on the users' journey to access data (Forster *et al.* 2025), on Member States' preparedness to implement it (Kessissoglou *et al.* 2024).or on the geopolitical dimension of data access under the EEDS (Donia & Marelli 2025).

Against this backdrop, from an methodological point of view, we examined the parliamentary debates (plenary discussions) and parliamentary work publications (MPs' and Senators' and institutional reports on bills) dedicated to the 2016 and 2019 health data access reforms. We also paid attention to the recent case law on health data access for research in order to get an understanding of and compare how judges refer to Privacy-Enhancing Technologies and balance it the public interest nature of such access. This discourse analysis demonstrates that within the 2016 and 2019 health data access reforms (**part 2**), Privacy Enhancing-Technologies were strongly promoted as the solution for health data circulation for research, to the point that, it was even anticipated in 2016 that health data would be made available in an open access format thanks to

[6] Few recent examples: Conseil d'État (CE), 25th April 2025 n°503163, CE 13th November 2024 n°492895; CE 19th October 2024 n°491644; CE 22nd March 2024 n°492369; CE 9 March 2023 n° 468007; CE 23rd November 2022,n° 456162.

protective Privacy Enhancing-Technologies (**part 3**). However, another key condition was also established and further elaborated in the following years: the requirement of public interest under which health data would be shared for research only if it benefited public interest such as for research purposes. As a result, Privacy Enhancing-Technologies and public interest narratives remain the framework guiding health data circulation for research in France. When privacy risk changed in nature, public interest discourses were used in the case law to authorise research regardless, thereby calling into question PETs as regulatory optimum (**part 4**).

2 The 2016 and 2019 French Health Data Access Reforms

In 2016, the French parliament identified the need to make health data available in open access to better inform the population about health and to make public decision. Such publication required to anonymised health data in order to protect patient's privacy (**2.1**). Anonymisation refers to methods that allow for the complete and irreversible de-identification of personal data, through the removal of direct and indirect identifiable information in data, or by using more sophisticated and often combined technical measures such as data synthesis, differential privacy and other obfuscation tools (for instance Sella *et al.* 2025). Under Privacy laws, anonymised data is no longer considered personal data and is therefore not subject to protection requirements. In comparison, pseudonymisation technics offers less privacy protections, removing identifier but leaving the data potentially open for re-identification. The 2016 Statute, therefore, introduced a dual system where some data would be available in open access as anonymised data while the remaining one would be accessible after pseudonymisation within an Information System (**2.2**). The 2019 Statute, on its part, expanded even more the System and reformed its governance (**2.3**).

2.1 Health Data: The Need for Open Access

In the 2010s in France, the need to access health data for public information and research purposes became paramount. Reports were solicited by the Government from various actors to pave the way for a health data circulation plan, and open access emerged to be the preferred option to do so. The Open data in Health Commission's report, published in 2014, defined open data as the openness and sharing of data published online in open format allowing for unrestricted and free re-use by anyone[7].To justify the need to organise such an open access in health, the report underlined that the former procedure for data access was too complex and lacked clarity for researchers, which was detrimental to

[7] Rapport Commission Open Data (2014), Ministère de la santé, p. 9 (https://drees.social-sante.gouv.fr/IMG/pdf/rapport_final_commission_open_data-2.pdf).

public health. On the other hand, the report indicated that the gathering of anonymised health data and its processing would constitute a source of progress for knowledge and value creation (p.37). Moreover, the publication in an open format would contribute to build a stronger health democracy by enabling patients to access data about the healthcare system and thereby empowering them to make more efficient decisions.

With those positive arguments, one can almost forget that a crucial factor presiding the adoption of the 2016 Act was also the tragic Mediator scandal in France. In the years prior to the reform, the use of the Mediator drug from Servier Pharmaceuticals for weight loss purposes, which was outside of its marketing authorisation, led to numerous patients' death or health issues that chocked the nation (Roure 2012) and abroad (Chrisafis, 2013). In France, it prompted discussions about more effective whistleblowing procedures, but also on how complicated access to health databases, which would have been crucial to shed light on the issue and provide material evidence to patients and their families, actually was (see Brasselet 2018, p. 339). In response the French Government at the time proposed to introduce provisions before Parliament that would facilitate health data access through open access: the future 2016 Act.

2.2 Health Data: The 2016 Act

The 2016 Health Act is a very large legislative text that introduced crucial rules for health data access for research. The new Act provided that anonymised health data should be accessed in an open format. As a consequence, no licence nor authorisation would be necessary to access it. The new Act also created the National Health Data System (NHDS), a large and centralised information system (a collection of databases) that was considered valuable for research purposes. Indeed, the NHDS encompasses the National Health Insurance database, which compiles all data on care reimbursement whether related to medical acts or prescribed drugs. This very large database contains data on more than 65 million patients (Moulis 2015)! However, the NHDS does not stop here, as it also includes the database dedicated to hospital activities and the national register of deaths for all French citizens. Under the new Act, such data would be made available for research purposes once pseudonymised and under the governance of a new institution: the National Health Data Institute.

The French parliament determined that six purposes could justify access to such a 'national treasure', as it now often referred to (Sénat 2023, Belot 2020):

- to provide information on health, healthcare provision, social care and on their quality;
- for the definition, implementation and evaluation of health and social protection policies;
- in order to understand health expenditures;

- to inform healthcare and social care professional on their activities;
- for health surveillance, monitoring and for safety;
- for research, studies, evaluation and for innovation in the field of health and social care[8] .

From a legal perspective, the evolution of the rules governing health data access - that is personal data (as opposed to anonymised data) - was worth analysing. However, the open access provisions of the Act remained largely unexamined. More generally, legal scholars, at this time, did not investigate the use of non-personal data for research considering, perhaps, that there was nothing to report on the subject. This, perhaps, could be explained by the anticipated application of another text adopted in 2016: the General Data Protection Regulation from the European Union. This much-awaited text was to be articulated with the 2016 Health Act on Health Data Access which drew considerable the attention in the literature. Such a coverage of the 2016 Act likely had the effect of diverting attention from Privacy-Enhancing Technologies, especially anonymisation techniques. Such a disregard for Privacy-Enhancing Technologies by the legal literature took also place in the analysis of the 2019 Health Act that also covered health data access for research.

2.3 Health Data : The 2019 Act

In 2019, the 2016 Act appeared to be underperforming. Indeed, the new NHDS did not function as smoothly as anticipated, and data access procedure remained lengthy for pseudonymised data[9]. To gain efficiency, the 2019 Act made changes on the governance side of the System in order to create the *Health Data Hub*, an agency designed not only to deliver practical access to the data but also to provide support for researchers seeking to use the System. The *Hub* emerged as the centre piece of the reform, its name in English chosen to attract researchers from around the world. However, the new governance was not the only innovation. Indeed, the 2019 Act also expanded the System, adding more databases to the list of available resources. Article L. 1461-1 of the French Public Health Code now provides that merely all data collected for a healthcare acts reimbursed by the French National Health Insurance Systems could be added to the NHDS. What an expansion!

The 2019 Act is more prolific regarding Privacy-Enhancing Technologies and emphasised that health data from the NHDS should be made available once pseudonymised as detailed in a security framework (referential) for personal data

[8] Article L. 1461-1 French public Health Code in its 2016 wording.

[9] Étude d'impact, Projet de loi relatif à l'organisation et à la transformation du système de santé, 13th February 2019, p. 89.

protection[10]. According to the Act, the main feature of this techno-legal framework should include pseudonymisation elements and access traceability measures. In its opinion on draft of the 2019 Act, the French Data Protection Authority, CNIL, cautiously indicated that Privacy-Enhancing Techniques should be sufficiently robust to protect such a large gathering of data (CNIL, 2019). While those elements of the text are central to protect privacy, they were, again, left out of most academic analysis or only briefly mentioned (Supiot 2020, Robin 2021) as a mere technical information. It must be conceded that the term 'Privacy-Enhancing technologies' had not yet emerged in 2016 and 2019 as the crucial notion it is now for describing tools and approaches aimed at preserving privacy. In this regard, the OECD report on Privacy Enhancing-Technologies, published in 2023, marked a turning point in public understanding of privacy solutions, therefore the term was not mentioned by the legal literature at the time, but the technics, mentioned in the Acts and framework, were not dissected even though the privacy discussion was crucial for securing the adoption of those Acts, as privacy concerns were central to the Parliamentary debates.

3 Privacy Enhancing-Technologies for Health Data Sharing: a Crucial Discussion

What role did Privacy Enhancing-Technologies play in the passage of 2016 and 2019 Acts? The analysis shows that anonymisation was central in supporting open access and facilitating the adoption of the 2016 Act (**3.1**), while also raising questions about the role that the Data Protection Authority should have in this scenario (**3.2**). Conversely, pseudonymisation was crucial for the circulation of identifiable data (**3.3**). However, Parliamentary debates revealed confusions regarding the distinction between anonymisation and pseudonymisation, leading to the impression that Privacy Enhancing-Technologies were also used to advance an agenda favouring data circulation rather than being critically assessed for their actual capabilities and limitations (**3.4**).

3.1 Anonymisation and the Open Access Discussion in Reports

Chapter 5 of the 2016 Health Act was dedicated to health data access and, interestingly, was entitled 'Enabling health data open access'. Very few scholars have noticed the significance of this title (with the exception of Robin 2021 or Péchillon 2015). In the wake – and in the shadow - of the Mediator scandal, 'health data open access' was presented as the Government's new open-data venture following the publication of various other public databases. Under the 2016 Act, article L. 1461-2 of the Public Health Code

[10] Loi n° 2019-774 du 24 juillet 2019 relative à l'organisation et à la transformation du système de santé, article 41.

provided, first of all, that the data from the NHDS would be made available for the public as aggregated data or in a way that prevents the direct or indirect identification of individuals. In other words, only anonymised data (therefore non-personal data) would be made available to the public from the newly created system. Such an option was considered feasible.

In the draft presentation of the 2016 Act before Parliament, the Health Minister (in charge of the text) was clear: ' We are allowing our nation to join the broad open data movement. It is our duty to promote health data for the collective good in the strict respect of privacy'.[11] In response, one Member of Parliament remarked that the draft did not demonstrate a robust technical oversights of health data open access.[12] When the question arose regarding the type of actors that could access the data, the answer was clear: no one would access identifiable data. Therefore, the technical protection of health data became an argument within the other crucial debate: determining who should have access to this data.

The 2013 *Bras Report* dedicated to health data, on which the provisions were based, was, indeed, optimistic that anonymisation techniques could be robust enough to ensure privacy, as the risk of reidentification would be residual (Bras and Loth 2015, p.29). The report indicated, at the time, that it would be very difficult for researchers to extract anonymous data and then compare them with other datasets in order to re-identify individuals (p.28). That particular risk was even judged 'less foreseeable' (p.28). However, the report, as well as the 2016 Act, remained unclear regarding what type of data could be shared openly with a limited re-identification risk. The 2014 *Health Data Report* attempted to provide clarity by listing databases that should be made open. The list included the swimming-pools water quality database and the drug consumption in hospital database (in volume), but it also listed the single parent benefit database (p.45). While the two first quoted databases contain non-personal data, the third one does and therefore would require the use of anonymisation tools. The risk of listing very different databases that are subjected to different privacy risk created much uncertainty on what privacy risks' exposure encompassed.

3.2 Anonymisation and the Data Protection Authority's Role

The parliamentary discussions centred on the techniques that could be used to make health data accessible in an open format. Report number 233 from the Parliamentary Joint Commission (a commission composed of MPs and Senators in cased of disagreement between the two chambers on a bill) evidenced these discussions and

[11] Translation by the author. Marisol Touraine, audition AN, 17 mars 2015 https://www.assemblee-nationale.fr/14/cr-soc/14-15/c1415034.asp

[12] Jean-Pierre Door, AN, 17 mars 2015 https://www.assemblee-nationale.fr/14/cr-soc/14-15/c1415034.asp

stated the necessity to have the French Data Protection Authority (CNIL) to provide clarity by listing robust anonymous technics that could be used to protect health data for open access: «The most sensitive personal data will only be accessible in open data after the application of complete anonymisation procedures declared compliant by the Data Protection Authority (CNIL), and any infringement of the ban on using open data in health for the purposes of identifying an individual may be subject to sanctions» (Translation by the author, CMP 2015).

'The most sensitive personal data' is a concept that contributed to blurring the line regarding the type of data being processed, given that health data are inherently sensitive by nature (without graduation). While the option of a CNIL seal of approval was contemplated by some MPs, it was later rejected as too complex to implement, since no single methodology existed for that purpose (CMP 2015). Despite this difficulty, the Government reassured MPs by suggesting that a later bill would introduce the possibility for the CNIL to list reliable and robust anonymisation techniques. This promising idea of the possibility to establish a list of sound and operational anonymisation idea of mandating the CNIL to list anonymisation technics did not materialise and the Government never brought it back again.

From the reports paving the way to the 2016 Act and its parliamentary discussions, we can conclude that, overall, Open Health Data was pursued at a time where health-related data - whether personal or not - was not easily accessible to the for informational purposes or to researchers. Indeed, in this context it felt paramount to ensure data access for all to facilitate a better understanding of health and healthcare related issues, both for citizens and public decision-making. However, by not sufficiently distinguishing between databases that included personal data and those already composed of statistical or aggregated data, the law created uncertainty regarding on the frontiers of open health data and the feasibility of anonymisation.

3.3 Pseudonymised Data and the Circulation Discussion

The 2016 Health Act provided that when data could not be anonymised, then it was to be made available under another Privacy-Enhancing Technologies tool: pseudonymisation. This was intended to be a residual measure, as the parliamentary reports on the Act demonstrated: 'When health data cannot be completely anonymised, their access for research, studies and public interest evaluation, will be limited and regulated by privacies guaranties' (Commission des Affaires Sociales 2015). The then-adopted article L. 1461-4.-I of the public health code stated that 'the NHDS does not provide individuals' first and family names or their social security number nor their address'. In order to enhance privacy further, the same article indicated that the medical practitioners' own identification numbers would be kept in a separate database. Indeed, Members of Parliament considered it too risky to keep all data together: the larger the dataset, the easier reidentification becomes.

In 2017, the Framework applicable to make data available for research in a pseudonymised form was adopted as a Ministerial Order[13]. The document provided that pseudonymisation should be irreversible and based on robust cryptographic techniques. Two rounds of pseudonymisation should take place, the first pseudonymisation should occur when data was transferred to the NHDS (level 1 pseudonymisation) and another round of pseudonymisation when the data was made available for researchers (level 2 pseudonymisation). Moreover, all accesses to data were to be logged in an access journal in order to identify who access it and when.

Despite privacy protection being adopted and enforced, the 2019 Act' s discussions demonstrated that privacy issues remained of importance. Indeed, while health data sharing for research was no longer questioned as such, the ways to protect privacy remained central to the new Act. Agnès Buzin, the then Health Minister, indicated during the debate that the most rigorous protection for Health Data was considered paramount by MPs.[14] In front of the Senate, she further indicated that access to data should be made through privacy-preserving approaches.[15] By emphasising on technical requirements for data sharing, the Minister did not question the necessity of data sharing, nor did she challenge the rights that individuals should retain over their data.

Even more privacy protections were added to the 2019 Act in order to prevent re-identification. Indeed, while in 2016, the law authorised to reidentify individual from the database if one could fear a significant health risk, to enrol the person in a trial or because it was necessary for the study (former article L. 1461-4 CSP), the 2019 Act suppressed this possibility to ensure that the pseudonymisation process remained irreversible.[16] The discussion, nevertheless, lacked clarity, the Health Minister indicating that the data from the HDH was now anonymised (Senate hearing 6, June 2019) even though it was actually pseudonymised, with a stronger prohibition on reidentifying patients.

However, the 2019 debate shifted focus away from open data to concentrate on access to identifiable data. Indeed, in practice, health research requires the most complete datasets possible in order to obtain sufficient elements to draw valid and useful conclusions. The trade-off between privacy protection and utility is very real and while open access health data is useful, the 4P approach to medicine (personalised, preventive, predictive, participatory) requires more than aggregated or statistical data. As a result, the way to ensure a smoother access to health data was addressed in 2019

[13] Translation by the author. Arrêté du 22 mars 2017 relatif au référentiel de sécurité applicable au Système national des données de santé

[14] Agnès Buzin, Assemblée Nationale, 18th March 2019

[15] Agnès Buzin, Sénat, 6th June 2019, https://www.senat.fr/seances/s201906/s20190606/s20190606_mono.html .

[16] Étude d'impact, Projet de loi relatif à l'organisation et à la transformation du système de santé, 13th February 2019, p.92.

without challenging the principles provided by the 2016 Act that – namely, data secrecy, confidentiality, traceability of access, and public interest reason to access data, the latter often being weighed against Privacy Enhancing-Technologies protections.

4 Privacy Enhancing-Technologies and Public Interest: a Dynamic Duo

The literature on the 2016 and 2019 Health Acts also emphasised that research projects will need to tick the 'public interest' box in order to gain access to the data. However, the notion is hardly defined by law (**4.1**) and is therefore left to interpretation by agencies and courts (**4.2**). Courts have, indeed, played a crucial interpretative role by establishing a balance that research project shall achieve between privacy protection and public interest, with the presence of the latter being enough to authorise research project when privacy protection were in doubt. Such a practical application of the statutory provisions quite differs from what Parliament foreseen in 2016 and 2019, namely, having two strong criteria that would not be watered-down by the other one (**4.3**).

4.1 The Public Interest Criterion in the 2016 and 2019 Acts

The 2016 Act introduced that when pseudonymised data is to be accessed, the access proposal should be justified by a 'public interest' dimension. In the parliamentary discussion in front of the Lower House (*Assemblée Nationale*), this 'public interest' criterion was deemed crucial when access was to be granted for identifiable data (*nominative* in French).[17] Therefore, the 2016 Act introduced a balance between privacy risk and protections on one hand, and the public interest of researches, on the other. Such a safeguard was intended to protect against undue access to health data, for instance by insurance companies seeking to increase premium or by other companies aiming to target advertising in the healthcare domain. However, public interest and what it entails is not defined or elaborated in the law, and, as others have noted public interest is a hard concept to grasp (Morlet-Haïdara 2022). One way to recompose its content is to examined the details in the French Public Health Code on what type of research could be made using the System. In this regard, Article L. 1461-1 III states that research project that will be allowed to access the NHDS must contribute to:

'1° The Information on health, healthcare provision, medical and social care and their quality;

2° The definition, implementation and evaluation of health and social protection policies;

[17] Parliamentary debates 19th march 2015 https://www.assemblee-nationale.fr/14/cr-soc/14-15/c1415040.asp .

3° Keeping track of healthcare expenditure, health insurance expenditure and medico-social expenditure;

4° The Information of health and medico-social professionals, structures and establishments about their activities;

5° Health surveillance, monitoring and safety;

6° To research, studies, evaluation and innovation for health and medico-social care'.

We can argue that these six borad purposes together give substance to the notion of public interest and clarify what is expected even though, in the law, the public interest criterion is an additional requirement to this list.

Article L. 1461-1V of the Public Health Code guides us further in defining the frontier of the 'public interest' criterion as it specifies the instances where access to the NHDS is prohibited. Indeed, access will not be granted to the System when the intended research seeks to:

'1° Promote medical products to health professionals or health establishments;

2° Exclude cover under insurance contracts or changes to insurance premiums for an individual or a group of individuals presenting the same risk.'

From the Parliamentary debate and the provision adopted in the Public Health Code, we can conclude that the public interest criterion does not, *de facto,* exclude private actors from the System, such as pharmaceutical or insurance companies, but rather require them not to use the System in a way that would negatively affect individuals or groups of individuals. The notion of public interest therefore aligns with a form of common interest that does not conflict with individual interest.

However, this criterion still lacks content and required both the HDH in charge of assessing proposals to access the NHDS, but also the Data Protection Authority (CNIL) and Courts to interpret it.

4.2 The Interpretation of the Public Interest Criterion.

The practice of assessing data access proposals led different actors to give life to the public interest criterion set out in the law. For the Health Data Hub (HDH), the notion of public interest should be understood as requiring research proposals submitted to tick the following boxes:

- « The aim of the project should be clear, intelligible and truthful
- The benefit of the project should be direct/indirect for groups of individuals, the society or the scientific community (bettering the healthcare system, research, increase of knowledge, etc.)

- Effort should be made toward transparency and result's publication of results, but also documentation, software used and link to public repositories
- Step should be taken to ensure scientific integrity measures, the quality of studies and prevent bias in results, to ensure the implication of research professionals, to put in place a proper scientific governance, to open [research] outcomes or methods in order to foster discussion and results' verifiability ».[18]

For the HDH, the public interest is therefore more about scientific integrity, transparency and openness, than about a precise field of research such as those described in the Public Health Code. The HDH goes further and specifies that the public interest criterion is not incompatible with commercial interests, underlining the fact that private actors can also access the System. However, the question of private actors' access to such data was not fully resolved and led the Data Protection Authority (CNIL) and the courts to,provide their own interpretation of the public interest notion.

Indeed, the CNIL and the *Conseil d'État* (the French highest administrative court) rejected access to the NHDS by journalists who wanted to publish an article on the list of 'best hospitals in France'. While in 2015, before the passing of the Act, the French Health Minister reassured that journalists would be able to access the NHDS for such studies as their work serves public interest[19], the CNIL did not felt as positive. For the CNIL and the *Conseil d'État* (that was later solicited with a demand to withdraw the CNIL's opinion), the methodology behind the data access proposal as well as the anticipated findings were questionable and lacking public interest.[20] According to the CNIL, the results would have negatively affected public information. In this light, we can conclude that the public interest criterion is assessed based on the scientific rigor of the methodology proposed to access the data. Accordingly, the CNIL and *Conseil d'État*'s decisions underline the need for a methodology strongly grounded within the scientific literature.

4.3 Public Interest *vs.* Privacy Enhancing-Technologies: a Question of Balance?

Since 2020, when it comes to public research, the CNIL and especially the *Conseil d'État*, both seem to weigh two different and, as demonstrated, equally important criteria: public interest and the use of Privacy Enhancing-Technologies, to the point where Daniel Kadar and his colleagues raised the following question: 'What if, in the Covid-19 era, health took precedence over the protection of health data?' (translation by the author, 2021) In other words, what if public interest in research were to trump privacy risk in the authorisation

[18] Translation by the author. HDH website : https://www.health-data-hub.fr/interet-public .

[19] Parliamentary hearing, 3th March 2015, https://www.assemblee-nationale.fr/14/cri/2014-2015/20150194.asp.

[20] CE, 30/06/2023, n° 469964, Sté d'exploitation de l'hebdomadaire Le Point; CNIL n° 2022-103, 20th October 2022.

of health data access? Such a question strongly emerged as the *Conseil d'État* in the above-mentioned case law[21] always authorises health data access when privacy doubts exist on the ground that the risk might only be residual while public interest into research commands to authorise access to data. The latest case law is an example of such findings and related to the creation of a 'super health database' for research purposes.

Indeed, in February 2025 the CNIL authorised the creation of a large database derived from the NHDS for the DARWIN EU project that is led by the European Medicines Agency for the EU. The project aims, in France, to create a database concerning 10 million patients in order to make the data available for the purpose of 'routinely estimate the prevalence and incidence of drugs and vaccines use in France using standardised indicators' and the prevalence or incidence of certain pathologies.[22] For the CNIL, such purposes satisfy the public interest criterion and the Privacy Enhancing-Technologies solutions applied appear robust: two level of pseudonymisation for the data and the erasure of the initial dataset within 3 months from its creation. However, the robustness of such protection was raised in front of the judges as data storage will be handled by an American Company: Microsoft.

For the HDH, in charge of the DARWIN EU project in France, Microsoft was the only available choice, no other French or European contractor being capable of storing such a large quantity of data. However, this choice raised concerns for charities dedicated to health as the company operates under Unites States of America (USA) Law. Indeed, since 2001, the USA laws authorised its federal agencies to access data held by their national companies. As a result, even if the data is stored in Europe, the US security agencies can still gain access to it, which would constitute an undue access for French citizens' personal data. For the plaintiffs (charities), the volume of data is so important that its double pseudonymisation would not be sufficient to protect the individuals' identities if federal US agencies were to access it.

For the judges and the CNIL, such a risk is deemed to be residual. The judges do not provide further details, but emphasised the public interest nature of the DARWIN EU project. Point 7 of the ruling states that there is a 'public interest in not delaying the carrying out of the studies relating to the estimation of the incidence and prevalence of pathologies in the general population planned as part of the DARWIN EU project' (translation by the author). Thus, the public interest prevails even when a risk of access exists, provided it is assessed as residual. In this regard, a previous decision of the *Conseil* d*'État* in 2024, on equivalent fact stated that 'it cannot be entirely ruled out that the authorised data processing, which is particularly sensitive in view of its nature as

[21] Conseil d'État (CE), 25th April 2025 n°503163, CE 13th November 2024 n°492895; CE 19th October 2024 n°491644; CE 22nd March 2024 n°492369; CE 9 March 2023 n° 468007; CE 23rd November 2022,n° 456162.

[22] CNIL (2025) : Deliberations n°2025-013 et 2025-014.

health data and of the scientific and economic potential of its use, may be the subject to access requests by US authorities, on the basis of the laws of that country, *via* the intermediary of the host's parent company'.[23] Therefore even when privacy risk remains, the public interest nature of proposed research positively influences the outcome of the ruling, raising questions about the original intentions of Parliament. Some might argue that privacy issues and public interest considerations are equally taken into consideration, but the rulings consistently appear to downplay privacy concerns. The reason might be perhaps because in 2016 and 2019, Parliament mostly anticipated private criminal ill-intent toward data access but not foreign State access to it and therefore courts are less able to tackle this issue or do not perceive it as a pressing one.

5 Conclusion

In 2016 and 2019 the French Parliament adopted new Acts that organised a System to make health data available for research. The study of the role played by Privacy-Enhancing Technologies in the parliamentary debate demonstrates that privacy concerns were very important for MPs. Indeed, health data were only to be shared under strong Privacy Enhancing-Technologies approaches such as robust anonymisation and double pseudonymisation schemes. Such protections were designed to prevent undue access to health data by private actors such as companies and illegal data re-identification of patients. The 2016 and 2019 Acts also introduced another criterion for data access: public interest. While this public interest remains evanescent by nature, it seems to refer both to the scientific quality of the research proposal and to its objective of benefiting the society.

In practice, the case law in front of courts demonstrates that the judges give considerable weight to the second element, public interest, over the first. Indeed, the limits of the Privacy Enhancing-Technologies are not thoroughly discussed even though they could be, as the cases often concern thousands and even millions of patients. Such a vast quantity of data could make re-identification technics and attacks easier. However, parliamentary debates had evidenced that privacy issues were imitatively linked to researchers and private companies' access. The case law shows that privacy risk is more complex, intertwined, and not restricted to private interests. Here foreign platforms such as Microsoft create new privacy risks, as their national government might have an interest in the data. Faced those new risks, courts are reluctant to engage in diplomatic rows and prefer focussing on the anticipated research's public interest in order to authorise their conduct especially since there is no national digital platform in France that has the material capacities to store such data volumes.

[23] Translation by the author. CE, n°491644, 19/10/ 2024

In order to address this new privacy challenge, the French Parliament adopted in 2024 a new Act that requires the NHDS and the HDH to use a French or and EU-based cloud for data storage within 18 months (art. 31 I)[24]. However, that deadline was postponed by the fact that some delegated legislation, that has not yet been published, are required to introduce more detailed provisions about the storage modalities. As a result, application of this new requirement has unfortunately been delayed leaving for a now an uncomfortable *status quo*.

[24] loi n°2024-449 visant à sécuriser et à réguler l'espace numérique.

References

Aufrechter, Cyril (2025): Le règlement européen des données de santé est publié ! In Dalloz actualité 20/05/25.

Belot, Laure (2020): Les données de santé, un trésor mondialement convoité. In Le Monde, March, 3rd 2020.

Bernelin, Margo (2024): Anonymisation des données et cybersécurité en santé : un droit hésitant ? in Droit, Santé et Société 67 (2), pp.20.

Bernelin, Margo (2020): Données massives et santé publique: entre redefinitions et ruptures normatives. In ADSP n°112, p.25-27.

Bernelin, Margo, Desmoulin, Sonia (2020): Données massives et santé publique. In ADSP n°112, p.2.

Bossi Malafosse, Jeanne (2016): Les nouvelles règles d'accès aux bases médico-administratives. In Dalloz IP/IT p.205.

Bossi Malafosse, Jeanne (2016a): La donneé de santé dans les systèmes d'information: du soin à la santé publique. In Communication Commerce électronique n° 10 étude 18.

Bras, Pierre-Louis, Loth, André (2014): Rapport sur la gouvernance et l'utilisation des données de santé, Drees.

Brasselet, Renatto (2018): La circulation de la donnée à caractère personnel relative à la santé: disponibilité de l'information et protection des droits de la personne, PhD Thesis, https://hal.science/tel-02188518v1.

Chazard, Emmanuel (2020): Big Data, data reuse en santé: un chemin semé d'embpuches nécessitant une approche pluridisciplinaire. In ADSP n°112, p.51-53.

Chrisafis, Angelique (2013): France shaken by fresh scandal over weight-loss drug linked to deaths. In The Guardian, January, 6th 2013.

CNIL (2019): Délibération n° 2019-008 portant avis sur un projet de loi relatif à l'organisation et à la transformation du système de santé (demande d'avis) n° 19001144.

Combes, Stéphanie (2022): Le Health Data Hub, levier pour la valorisation des données de santé. In Annales des Mines - Réalités industrielles, Août (3), pp. 59-62.

Commission Mixte Pariataire (CMP 2015), Rapport Projet de loi de modernisation de notre système de santé, n°233.

Commission des affaires sociales (2015), rapport ,°3215, 10th December 2015.

CybersecurityAsia.net (2024): Cyber Predators Target Vulnerable Victims: Hackers Blackmail Hospitals, Trade Patient Data and Find Partners Through Darknet Ads', sept. 25th 2024 (https://cybersecurityasia.net/hackers-target-healthcare-darknet-ads/).

Debiès, Elise (2016): L'ouverture et la réutilisation des données de santé: panorama et enjeux.In RDSS p.697.

Desmoulin-Canselier, Sonia (2018): L'évaluation des médicaments à l'ère de la médecine des données. In RDSS p.1043.

Devillier Nathalie (2017): Chapitre 6. Les dispositions de la loi de modernisation de notre système de santé relatives aux données de santé. In Journal international de bioéthique et d'éthique des sciences 28(3), pp. 57-61.

Donia, Joseph & Luca Marelli (2025): Anticipating ethical and social dimensions of the European Health Data Space: A rapid systematic review. In Health Policy, Volume 162,105443.

Forster Rachel *et al.* (2025): User journeys in cross-European secondary use of health data: insights ahead of the European Health Data Space, In *European Journal of Public Health*, Volume 35, Issue Supplement_3, pages iii18–iii24.

Garcia, anna Cristina Bicharra., Garcia, Marcio Gomes Pinto. & Rigobon, Roberto (2024): Algorithmic discrimination in the credit domain: what do we know about it?. In *AI & Soc*, n° 39, 2059–2098.

de Grove-Valdeyron, Nathalie (2024): Espace européen des données de santé: enjeux et défis pour l'utilisation secondaire des données de santé. Entre gouvernance des données et interlligence artificielle: quelle place pour la poursuite de l'intérêt général ?In Obavia, pp.18-23.

de Grove-Valdeyron, Nathalie ed. (2025): *Espace Européen des Données de Santé et IA. Toulouse*: Presses de l'Université Toulouse Capitol.

Horgan, Denis, *et al.* (2022): European Health Data Space—An Opportunity Now to Grasp the Future of Data-Driven Healthcare. In Healthcare 10, no. 9: 1629.

Hoeyer, Klaus, *et al.* (2024): Health in data space: Formative and experiential dimensions of cross-border health data sharing. In Big Data & Society, 11(1).

Kadar, Daniel, Abdesselam, Stéphanie, Gaillard, Laetitia (2021): Données de santé : un vecteur d'innovation sous trop haute surveillance ?. In La Revue des juristes de Sciences PO n° 21, p.10.

Kessissoglou, Irini *et al.* (2024): Are EU member states ready for the European Health Data Space? Lessons learnt on the secondary use of health data from the TEHDAS Joint Action. In European Journal of Public Health, Volume 34, Issue 6, pages 1102–1108.

Kleinman, Zoe (2020):, ‘Therapy patients blackmailed for cash after clinic data breach’, BBC.com, oct. 26th 2020 (https://www.bbc.com/news/technology-54692120).

Lianos, Ioannis (2025): Access to Health Data, Competition, and Regulatory Alternatives: Three Dimensions of Fairness. In Journal of Competition Law & Economics, nhaf016.

Marcus, J. Scott; Martens, Bertin; Carugati, Christophe; Bucher, Anne; Godlovitch, Ilsa (2022): The European Health Data Space. IPOL- Policy Department for Economic, Scientific and Quality of Life Policies, European Parliament Policy Department studies.

Marelli, Luca *et al.* (2021): Big Tech platforms in health research: Re-purposing big data governance in light of the General Data Protection Regulation’s research exemption’, In Big Data & Society, 8(1).

Marelli, Luca *et al.* (2023): The European health data space: Too big to succeed? In Health Policy V.135, 104861.

Megerlin, Francis (2024): Espace européen des données de santé : portée de la proposition de règlement. In Reccueil Dalloz 119.

Mercier, Sandra (2020): Médecine Génomique; vers une médecine predictive. In ADSP n°112, pp.30-32.

Morlet-Haïdara, Lydia (2018): Le système national des données de santé et le nouveau régime d’accès aux données. In RDSS p.91.

Morlet-Haïdara, Lydia (2022): Problématiques juridiques posées par le Big Data et les outils institutionnels de la recherche en santé. In Santé Publique, . 34(3), 335-344.

Moulis Guillaume *et al.* (2015): French health insurance databases: What interest for medical research?. In La Revue de Médecine Interne, Volume 36, Issue 6, 2015, pp. 411-417.

OECD (2023): Emerging Privacy-enhancing technologies Current regulatory and policy approaches.

Pailhès, Bertrand (2018): Comment définir et réguler les « données d’intérêt général » ?. In Annales des Mines - Enjeux numériques, 2(2) pp. 39-43.

Pechillon, Éric (2015): L’accès ouvert aux données de santé : la loi peut-elle garantir tous les risques de dérives dans l’utilisation de l’information ?. In L'information psychiatrique, 91(8), pp.645-649.

Quinn, Paul and Erika Ellyne, Cong Yao (2024): Will the GDPR Restrain Health Data Access Bodies Under the European Health Data Space (EHDS)?. In Computer Law & Security Review, Volume 54,105993.

Robin, Agnès (2021): Chapitre 3. Open data et santé : quelles modalités pour la diffusion et l'exploitation des données de santé ?. In Journal international de bioéthique et d'éthique des sciences 32(2) pp. 33-44.

Roure, Thomas (2012): L'affaire Mediator : retour sur 18 mois de scandale. In *Le Monde*, 14th May 2012.

Shabani, Mahsa (2022): Will the European Health Data Space change data sharing rules? In Science 375,1357-1359.

Sella, Nadir *et al.* (2025): Preserving information while respecting privacy through an information theoretic framework for synthetic health data generation. In npj Digit. Med. 8, 49.

Sénat (2023) : Données de santé, rapport d'information n° 873 (2022-2023).

Supiot, Elsa (2020): Du secret médical à la mise à disposition des données de santé - le Health data hub. In Revue des contrats, 112, p.94.

Tamba, Julie (2025): Interopérabilité des dossiers médicaux : ce qui change avec l'espace européen des données de santé. In Journal de droit de la santé et de l'assurance maladie, 43.

Teller, Marina (2022): La régulation des données de santé : entre intérêt général et intérêts particuliers Introduction au cahier spécial. In Revue internationale de droit économique, t.XXXVI(3) pp. 5-11.

Societal Impact of Digital Credentials on Vocational Training in Latin America

Alexander Nussbaumer[1], Carlos Alario-Hoyos[2], Carlos Delgado Kloos[2], Chiara Russ-Baumann[1], Carina Kern[1], Miguel Antonio Morales Chan[3], Hector R. Amado-Salvatierra[3], Luis Eduardo Veliz Argueta[4], Christian Gütl[1]

[1]Graz University of Technology, Austria

[2]Universidad Carlos III de Madrid, Spain

[3]Galileo University, Guatemala

[4]Fundacion Kinal, Guatemala

DOI 10.3217/978-3-99161-062-5-017, CC BY 4.0

Abstract. This paper presents initial results of a pilot project that aims to introduce and sustain the use of digital credentials in Latin America. Digital credentials represent a major innovation, supporting the modular and flexible learning paths necessary for continuous reskilling in today's fast-changing labor market. Latin America can even more benefit from digital credentials, as a vulnerable youth exists that faces heightened barriers to education, employment, and training. However, there are severe challenges to introducing digital credentials in vocational training especially in Latin America, due to legal uncertainties, lack of standardization, limited interoperability, a weak digital culture in institutions, and a fragmented situation of educational institutions. As a solution approach, we have set up a training and awareness programme that introduces key concepts to all relevant stakeholders in the vocational training area in Guatemala, accompanied by pilot implementation for institutional demonstrations. In a survey, young people in Guatemala reacted very positively to this aim of introducing digital credentials and an expert group outlined opportunities and pitfalls.

1 Introduction

In Europe, there is a growing trend towards the adoption of digital credentials (electronic certificates that verify an individual's acquisition of skills or knowledge), serving as digital equivalents of traditional paper-based credentials. While digital credentials require a technological infrastructure for issuance, storage, and verification, they offer clear advantages: enhanced portability, richer documentation of the learning process, and greater trust from employers through reliable, verifiable information (Grech et al., 2021).

Digital credentials represent a major innovation, supporting the modular and flexible learning paths necessary for continuous reskilling in today's fast-changing labor market. Several key European initiatives exemplify this shift, including the European Digital Credentials for Learning (EDC)[25], the European Blockchain Services Infrastructure (EBSI)[26], and the CertiDigital project in Spain[27]. The EU Digital Education Action Plan 2021-2027[28] has also played a vital role in addressing the challenges and opportunities of digital education in Europe. It promotes the development of digitally certified qualifications frameworks, such as the European Qualifications Framework (EQF) and the European Skills, Competences, Qualifications and Occupations (ESCO) framework. The Europass platform, which is compatible with EDC, allows for the issuance and exchange of multilingual, verifiable digital credentials across borders.

In Latin America, while some higher education and vocational institutions have begun experimenting with digital credentials, there is still no unified ecosystem, nor shared standards or legal frameworks to ensure data security, privacy, and interoperability. The lack of government commitment further highlights the need for awareness-building and capacity development within Vocational Education and Training (VET) institutions. To keep pace with global trends, VET institutions must establish robust digital credential ecosystems. Such systems would allow:

- Citizens to build online learning portfolios, apply for jobs or training opportunities using verified credentials, and retain lifelong control over their learning data.
- Training providers to issue standardized digital credentials at lower cost, ensure quality, and improve institutional mobility and credit recognition.
- Employers to quickly authenticate credentials, better understand applicants' competencies, and detect fraudulent or tampered documentation.

Digital credentials can carry rich metadata, including learner identity, training details, learning objectives, theoretical and practical workload, assessment methods, qualification level, and quality assurance procedures (Kemcha et al., 2024). These features enhance the credibility, transparency, and comparability of educational offerings across institutions and countries. Despite these benefits, early implementations in Europe have revealed challenges, including legal uncertainty, lack of standardization, limited interoperability, and a weak digital culture in institutions. Most VET centers

[25] https://europass.europa.eu/en/stakeholders/european-digital-credentials

[26] https://ec.europa.eu/digital-building-blocks/sites/display/EBSI

[27] https://certidigital.uc3m.es

[28] https://education.ec.europa.eu/focus-topics/digital-education/action-plan

currently lack the expertise and internal capacity to design and implement digital credential systems.

These challenges are even more pronounced in Latin America, where educational systems are more fragmented. For this reason, initiatives are critical for laying the foundation for regional digital credential ecosystems, promoting adoption, trust, and collaboration among key stakeholders. This paper reports a pilot project that aims to research and establish a concept, how digital credentials can be introduced in Latin America. After discussing related work on vocational training and digital credentials in the next section, the concept and plan of the pilot project is presented in Section 3. In Section 4 the feedback from different stakeholder groups are presented, which outlines the needs, opportunities, and challenges of introducing digital credentials in Latin America.

2 Related Work and Background

2.1 Vocational Training

In today's world of rapid digitalization, globalization, and climate-related transformation, economies increasingly require adaptable, skilled workers. Traditional education models alone are no longer sufficient to meet these shifting demands (Maclean & Lai, 2011). As a result, vocational education and training (VET) has gained global relevance. These programs promote lifelong learning and equip individuals with practical, occupation-specific competencies often in collaboration with industries (Todd & Dunbar, 2018). While once limited to manual trades and often undervalued, vocational education now spans diverse sectors and contributes significantly to employability (Miller, 2024). Around 80% of jobs worldwide require vocational skills, prompting many countries to invest in VET as a way to facilitate economic growth by strengthening employability, especially for marginalized groups (Maclean & Lai, 2011).

Depending on the geographical location, VET slightly differs in terms of its realization, and therefore the terminology also varies by region. For example, in Europe we speak of vocational education and training (VET), while in the USA career and technical education (CTE) has become established. However, all these expressions refer to the same field. VET can therefore be defined as 'education and training which aims to equip young people and adults with knowledge, skills and competences required in particular occupations or more broadly on the labour market' (Erasmus+ Programme Guide, n.d.)

Many national vocational training authorities (VTAs) aim to strengthen ties between training providers and industries, enhancing economic impact (Angélica Ducci, 1997). Because informal training is the most established form of vocational education, many

Latin American nations aim to recognize informal apprenticeships to improve certification accessibility (Suescún Barón et al., 2024).

In the Latin American context, the COVID-19 pandemic amplified pre-existing disparities, particularly among vulnerable youth who face heightened barriers to education, employment, and training. Although Technical and Vocational Education and Training (TVET) systems across the region have made efforts to integrate digital skills into their curricula, a significant mismatch persists between the availability of digital training and the inclusion of marginalized populations. Recent studies by the International Labour Organization (ILO) highlight that, while enrolment of vulnerable youth has increased, their sustained participation and success are often hindered by recruitment challenges, entry barriers, and insufficient institutional support (Morales et al., 2022). Countries across the region are implementing policies to modernize vocational training systems, strengthen quality assurance mechanisms, and incorporate digital and green skills aligned with future labor demands (Prada & Rucci, 2023; Hirsch 2025).

Guatemala provides a valuable case study in this regard. The Instituto Técnico de Capacitación y Productividad (INTECAP) stands out as a national institution that has modernized its vocational training model by incorporating digital competencies, industry-aligned certifications, and decentralized strategies to improve access across diverse territories. INTECAP's approach reflects a comprehensive model that combines technical excellence with responsiveness to labor market trends (Cinterfor/ILO, 2001). Nonetheless, the challenge of reaching Guatemala's most vulnerable youth—particularly those in rural and indigenous communities—remains significant and requires deliberate strategies and targeted support systems.

In parallel, Fundación Kinal exemplifies how non-governmental organizations can complement national efforts by offering vocational training rooted in ethical values, personal development, and social responsibility. Kinal's programs focus on youth from low-income backgrounds, providing technical training and job readiness skills in partnership with private sector stakeholders. Such institutions play a key role in filling gaps left by formal education systems, not only by promoting employability but also fostering social inclusion and upward mobility.

Across the region, the successful integration of vulnerable youth into digital and vocational education systems requires inclusive institutional mandates, the incorporation of 21st-century competencies into curricula, stronger teacher training initiatives, and enhanced inter-institutional coordination. As the ILO emphasizes, policy frameworks must evolve from broad-based access to more targeted interventions that reflect the lived realities of young people navigating economic precarity, informal labor markets, and limited digital access (Morales et al., 2022). The evolving nature of work in Latin America calls for integrating soft skills, digital literacy, and environmental awareness into vocational curricula (Salazar-Xirinachs & Vargas, 2017; Amado-Salvatierra, Morales-

Chan, Hernández-Rizzardini, 2024). As such, vocational education in the region is increasingly viewed not only as a means of immediate job placement, but as a cornerstone of lifelong learning and sustainable development.

2.2 Digital Credentials

Digital credentials are electronic representations of learning outcomes and achievements, encompassing a broad spectrum such as digital certificates, badges, micro-credentials, macro-credentials, and verifiable credentials (Kato et al., 2020; AACRAO, 2022). Despite ongoing inconsistencies in terminology and overlaps between related concepts (Brown et al., 2021; Keevy & Chakroun, 2015), the shared characteristic of digital credentials is their digital format, which allows for secure storage, efficient transmission, and automated verification (Brands, 2002). At their core, digital credentials represent a shift away from traditional paper-based documentation towards systems that promote trust, portability, and authenticity through digital technologies (Foshay & Hale, 2017). They mark an evolution in how qualifications, competencies, and skills are recorded and shared, offering interoperable and secure solutions that are increasingly applied across various sectors, including education and training (Quigley, 2023; UNESCO, 2022).

The technical ecosystem of digital credentials involves several key roles: issuers (such as universities or certification bodies), holders (individuals who receive and store credentials), and verifiers (institutions or employers who check their validity) (World Wide Web Consortium, 2024). A credential typically goes through a lifecycle starting with issuance, followed by storage, sharing, verification, and potentially revocation (Gräther et al., 2018; Sedlmeir et al., 2021). This process is underpinned by a set of technological foundations that make digital credentials reliable and secure (Mühle et al., 2023). At the heart of the digital credential infrastructure are cryptographic technologies. Digital signatures ensure that a credential has not been tampered with and confirm its origin (Katz & Lindell, 2007). Public Key Infrastructure (PKI) and asymmetric encryption allow secure communication and authentication of identity (European Union Agency for Cybersecurity, n.d.). Additionally, emerging technologies such as Zero-Knowledge Proofs enable privacy-preserving verification by allowing credential holders to prove claims without revealing all underlying data (Goldreich et al., 1986). Another critical concept is Self-Sovereign Identity (SSI), which enables individuals to control their own digital identities and manage their credentials independently, often using Decentralized Identifiers (DIDs) (Preukschat & Reed, 2021; World Wide Web Consortium, 2022a). Moreover, blockchain and other distributed ledger technologies (DLTs) play a central role in making credentials tamper-evident and publicly verifiable without relying on a central authority (Narayanan et al., 2016; Yli-Huumo et al., 2016; Grech & Camilleri, 2017).

Two major international standards structure the implementation of digital credentials: The Open Badges standard, developed initially by Mozilla in 2012 and now maintained by

1EdTech, and the W3C Verifiable Credentials (VC) standard, maintained by the World Wide Web Consortium (W3C). Open Badges define a portable digital credential, originally designed exclusively as a visual badge embedded with metadata about the issuer, recipient, learning outcomes, criteria, and evidence. These badges are used across various sectors for both formal and informal learning. Until version 2.1, Open Badges were typically encoded in JSON and often embedded in PNG image files for easy sharing and display on digital portfolios or social media. Over 40 million Open Badges have been issued globally, providing a standard way to represent and communicate skills and achievements (1EdTech, 2024; UNESCO, 2018). In contrast, the W3C Verifiable Credentials standard offers a more technical and decentralized model focused on secure, privacy-respecting credential exchange. A Verifiable Credential (VC) enables cryptographic proof of authenticity and integrity and supports Self-Sovereign Identity (SSI) frameworks. It allows for selective disclosure, giving the holder control over which parts of the credential are shared with a verifier. Each VC involves three roles: issuer, holder, and verifier, with verification performed through cryptographic proofs without requiring a central authority. While the use of blockchain is optional, it can support functions such as credential revocation and timestamping, therefore enhancing transparency (W3C, 2022b; 2024). VCs are increasingly adopted in high-stakes domains such as academic qualifications, identity verification, and professional certifications.

Originally, these two standards had different technical architectures and goals: Open Badges focused on motivating learners and recognizing achievements in accessible and visual formats, while Verifiable Credentials emphasized security, privacy, and interoperability for scalable digital trust ecosystems (Lemoie, 2024, 2025). However, with the release of Open Badges 3.0, this gap has been closed. Open Badges 3.0 is now fully compatible with the W3C Verifiable Credentials Data Model v2.0 and the previous version v1.1 (IMS Global, 2025). It retains the core structure and use cases of Open Badges (e.g., skills recognition, portable evidence of learning) but is implemented as a Verifiable Credential using JSON-LD, Decentralized Identifiers (DIDs), and cryptographic proofs (IMS Global, 2025; Lemoie, 2024, 2025).The VC Data Model thereby provides the structure for every process in the lifespan of a VC, whereas credential-type-specific schemas function as standardized templates that define how particular types of achievements are represented and understood within that structure (IMS Global, 2025). Therefore, the convergence of both standards makes it possible to have a unified, open, and trustworthy digital credentialing ecosystem, where the respective advantages are combined.

The development and implementation of digital credentials are advancing rapidly worldwide. In Europe, several strategic frameworks and infrastructures have been established. Notable among them is the European Digital Credentials for Learning (EDC) system, which builds on the Europass initiative to create interoperable and trusted formats for issuing and verifying digital diplomas and certificates across borders

(European Commission, 2020, 2023). The European Blockchain Services Infrastructure (EBSI) provides a decentralized framework for secure credential exchange, while legislation like the GDPR and the eIDAS Regulation ensures data protection and legal validity for digital signatures and trust services (EUR-Lex, 2016; European Commission, 2024).

In Latin America, developments are more decentralized but show strong momentum. Universities in countries like Mexico, Chile, Argentina, and Peru are issuing digital credentials for academic and extracurricular achievements (Tecnológico de Monterrey, 2023; Universidad de Chile, n.d.; Pontificia Universidad Católica del Perú, 2024). Additionally, national vocational training institutions in countries such as Brazil and Colombia have implemented blockchain-based credentialing systems (SENAI, Confederação Nacional da Indústria, n.d.; SENA, 100.000 Strong in the Americas, 2024). The Inter-American Development Bank (IDB) is also a key driver in the region, promoting digital badges and micro-credentials for upskilling and regional mobility (Porto & Presant, 2023).

Despite the rapid progress and the many opportunities digital credentials offer, several challenges remain. Privacy and data security must be carefully managed, particularly when sensitive personal data is involved. Moreover, broad adoption across sectors and regions will require trust in the issuing authorities, the technical systems used, and the long-term viability of the infrastructures (Mühle et al., 2023). Digital credentials represent a transformative development in how learning, skills, and qualifications are documented, recognized, and exchanged. The combination of technological innovation, legal frameworks, and institutional collaboration is paving the way for more equitable, efficient, and learner-centered education and employment systems worldwide.

2.3 Role of AI in the Context of Learning Content and Digital Credentials

Generative Artificial intelligence is impacting all aspects of our life and work. Does it also play a role in the context of microcredential courses and digital credentials? Let us define a framework for the design and deployment of microcredentials covering from specification to learning opportunity and to certification (see Fig. 1).

Figure 1: Microcredential Framework with one learning opportunity and certification

The specification describes the knowledge and skills that are going to be transmitted in the course, together with a description of activities and assessment formats. The specification is useful for course catalogues. A learning opportunity is a concrete

instantiation of this specification that takes place at a particular time and place with concrete teachers, students, and learning material, activities, and assessments. The certification is a claim of the achievements of one particular learner in one opportunity. There is one-to-many relationship among these concepts: the same specification can be used for several learning opportunities, and one opportunity implies the issuing of several certificates, one for each learner describing their achievements (see Fig. 2).

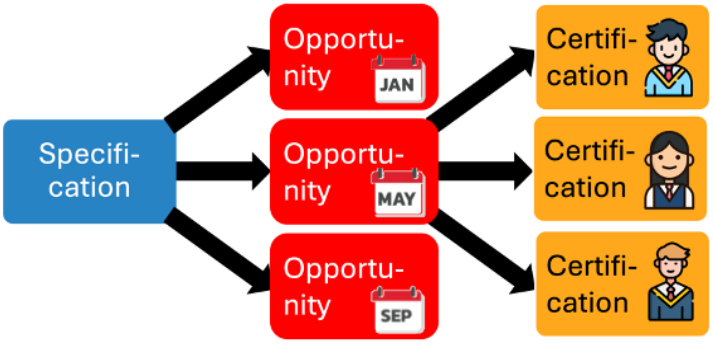


Figure 2: Microcredential Framework with multiple learning opportunities and certifications

Let's dive into the details of the opportunity. The educational material has to be prepared following the specification. This might include the preparation of text documents, videos, podcasts, pictograms, mind-maps, exams, and more. Then, the enactment comes with the teacher teaching the class and the learner following the class (which could be in presence or online). Finally, the learner must do some work, individually or in groups, to assimilate the content and do the assessments.

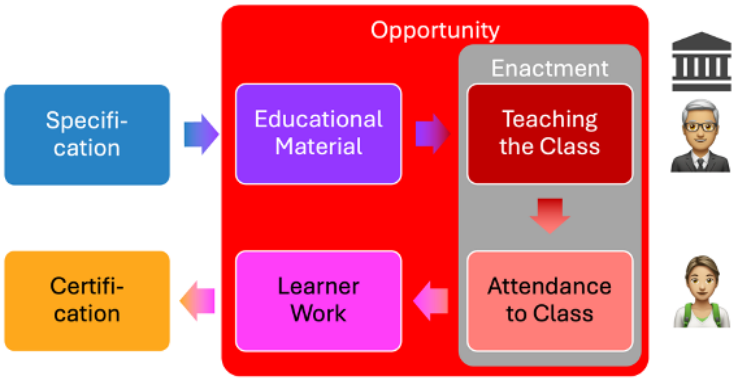


Fig.3: The learning opportunity process in detail

Once this framework is defined, the question is where can we apply Artificial Intelligence? The short answer is everywhere. Let's give some examples of the role AI can play in each of these 6 boxes. More information can be found in Delgado Kloos et al. (2025).

Specification. There are several classifications for defining skills and competences. For example, ESCO (esco.ec.europa.eu) defines close to 14,000 skills. Lightcast

(lightcast.io/open-skills) has also a rich taxonomy with more than 33,000 skills. AI can help in providing its first proposal for filling out a specification.

Educational Material. It is well-known that AI can help to prepare material in all formats: text, images, audio, video, etc. The technology has advanced so much that it is even possible to create videos with an avatar of teachers or translate their audio to other languages with voice cloning and lip synchronization. See for instance the MOOC 'Insegnare con l'AI: Strumenti'[29], which was recorded in Spanish but is shown in Italian (Fig. 4).

Fig.4: MOOC 'Insegnare con l'AI: Strumenti' that has been translated to Italian language.

Teaching the Class. When teaching a class, it is necessary to orchestrate activities to achieve maximum engagement of the students. AI tools can give useful recommendations building upon the rich literature in education and pedagogy.

Attendance in Class. Specific applications like StudyFetch (studyfetch.com) can help in transcribing the live lecture of the teacher so that the learner can later ask specific questions for clarification or even request quizzes on the material explained.

Learner Work. Multiple bots have been developed that help tutor learners on specific courses, such as Khamigo (khanmigo.ai), Duolingo Max (en.duolingo.com/help/what-is-duolingo-max), or CharlieBot at UC3M (uc3m.es/sdic/en/servicios/charliebot). But without going so far with specific developments that use RAG (Retrieval-Augmented Generation), it is possible to use general-purpose AI tools such as NotebookLM (notebooklm.google) and upload the educational material (PDFs, websites, Youtube videos, etc.) to specialize the responses to this material. Student forums such as David Malan's AI Duck (Liu et al. 2025) help students by adding an AI to the participants of the forum.

Certification. Certification or credentialing is closely related to specification. Both describe the learning opportunity in an abstract way, the specification focusing on the

[29] http://federica.eu/esplorare-ia

overall learning goals (or intended learning outcomes) for all and the certification on the concrete learning outcomes achieved by each learner. To achieve interoperability across systems, digital credentials are coded into particular formats, such as EDC (European Digital Credentials) in the context of ELM (European Learning Model) defined by the European Commission[30] or OpenBadges as maintained by 1EdTech (1EdTech, 2024). We know that AI is becoming increasingly better at coding. Therefore, the role of AI in digital credentialing is not only in helping fill out the necessary fields, but also in presenting them neatly in tabular form for human inspection or in one of the digital formats for interoperability (Delgado Kloos et al., 2025).

We have seen that there are multiple places in the context of microcredentials where AI can play a supporting role. However, one should use AI thoughtfully. There are many dangers and issues which are still unresolved.

3 Erasmus+ Project EcoCredGT

Digital Credentials are increasingly viewed as valuable tools for recognizing learning and competencies in formal and non-formal education contexts. From the perspective of educational researchers, digital credentials are seen as mechanisms to help learners evidence specific skills, reflect on personal development, and gain a competitive advantage in the labor market (Miller et al., 2017). This section presents the Erasmus+ project EcoCredGT (ecocredgt.org) that aims to set up a training and awareness programme introducing key concepts to all relevant stakeholders in the vocational training area in Guatemala, accompanied by pilot implementation for institutional demonstrations. The training concept and pilot implementation includes the whole cycle of creating learning content that is deployed as training units, followed by issuing digital credentials for mastering the units. Details on digital credentials, creation of training units, and vocational training as explained in the last section serve as foundation for this course. Before describing the training concept, the overall aim of the project is presented in the next subsection.

3.1 Project Goals

The project aims to build capacity in Vocational Education and Training (VET) institutions by fostering a digital credentials ecosystem that positively impacts employability. This ambitious global goal is broken down into five specific objectives:

1. To strengthen the capacities of VET institutions in driving digital transformation.

[30] http://europass.europa.eu/en/stakeholders/european-digital-credentials

2. To develop a replicable model for a digital credential issuance center that can be adopted by other VET institutions.
3. To implement pilot projects for vocational and professional training micro-courses that enhance employability, issuing the corresponding digital credentials for those learners who complete these courses.
4. To create a stronger connection between VET institutions and society by raising awareness of the value of digital credentials and updating training programs to meet the challenges of the Fourth Industrial Revolution.
5. To establish an observatory of digital credential issuance centers that fosters a community for sharing best practices, success stories, and training across the region.

The Erasmus+ project is implemented in Guatemala, with the leadership of two educational institutions: Fundación Kinal and Universidad Galileo. The project also includes two European partners with strong research experience in the field of educational technology: Universidad Carlos III de Madrid (Spain) and Graz University of Technology (Austria). The model for a digital credential issuance center is expected to be replicated in other countries in the Latin American region and to raise awareness of the need for the adoption of digital credentials in the region through the observatory.

3.2 Project Implementation

Based on the goals listed above and aligned with the overarching vision, the project conducted a comprehensive exploration of the global landscape of the digital credential scene and has carried out a detailed self-diagnostic analysis within its partner institutions in Guatemala. This study has not only mapped international best practices but has also identified institutional strengths and capacity gaps among the participating VET institutions.

These efforts have been fundamental in defining the roadmap for the design and implementation of a comprehensive Multilevel Training Plan, carefully tailored to the specific profiles and needs of the diverse actors involved in the digital credential ecosystem. The program targets four primary groups: (a) administrative and technical staff, who are central the operational deployment of credentialing platforms and require training in standards such as Open Badges and digital issuance tools; (b) educators, who must master the design of learning outcomes, assessment criteria, and credential metadata to ensure meaningful certification of competencies; (c) students, as the primary beneficiaries, who will engage with courses that certify transversal and technical skills enhancing their employability; and employers, whose awareness and endorsement are critical for the labor market acceptance of credentials, and who are invited to integrate them into hiring processes. By identifying institutional needs, contextual constraints, and

global benchmarks, the project has developed a differentiated training strategy that ensures relevance, accessibility, and scalability for each stakeholder group.

The Multilevel Training Plan is structured in three main modalities: (a) Webinars, (b) Massive Open Online Course (MOOC), and Hybrid workshops. This level approach (see figure 1) ensures that participants not only acquire theoretical knowledge but also develop the practical competencies necessary for implementing and sustaining digital credentialing processes within their respective institutions and professional environments.

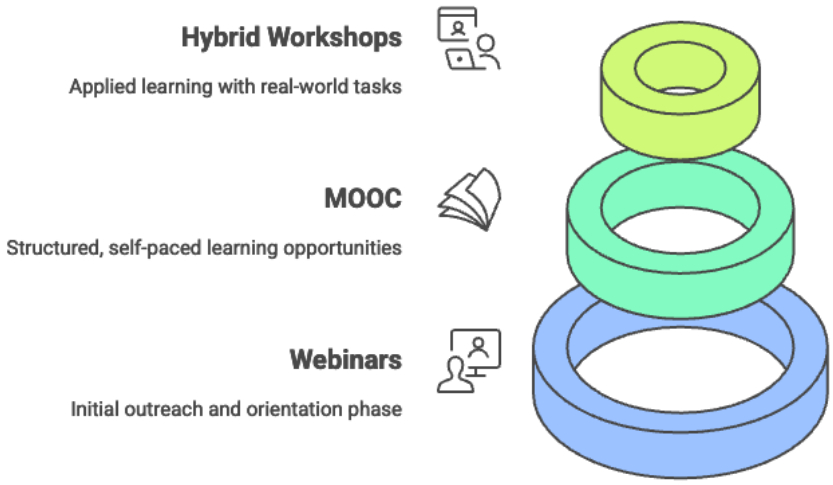


Fig.5: Multilevel Training Plan targeting four primary groups

(a) The Webinar series: composed of four executive format sessions serves as the initial outreach and orientation phase. These two-hour events are strategically designed for high accessibility and minimal time commitment, allowing a broad audience to quickly engage with core ideas. Each session targets a distinct stakeholder group: educators, technical and administrative staff, and employers. The first webinar provides from the introduction to digital credentials, covering key topics such as the evolution from badges to certificates, the transition from OpenBadges to the European Learning Model, the fundamentals of micro-credentials, and examples from global implementations. This second webinar will continue the work started in the State of the Art analysis. We will explain the technical ideas behind digital credentials in a simple way. The session will also show how credentials are created and shared, using real examples to help participants understand their practical use. In this third webinar, we will explore two main approaches to implementing credentialing systems: in-house implementation, where institutions use their own infrastructure to manage the process; and the use of external service providers, which offers benefits such as interoperability, scalability, and alignment with international standards. In the last webinar, we will discuss employers' perspectives on digital credentials, the potential benefits they see in adopting these systems, and what

this means for job seekers and students. The session will highlight how digital credentials can enhance employability, improve talent matching, and support lifelong learning.

(b) The MOOC: expands on these foundations, offering structured, self-paced learning opportunities designed to deepen and consolidate the knowledge introduced in the webinar series. Our MOOC will follow a modular structure, with weekly units composed of short instructional videos, learning activities, and formative assessments. Each unit will include clearly defined learning objectives and culminate in the production of applied learning activities and a final integrative project, aligned with internationally recognized credentialing standards.

Considering that the success of a MOOC is largely dependent on the level of student engagement and sustained participation (Hernandez et al., 2014), we have integrated the use of artificial intelligence tools to enhance the instructional design and foster a more meaningful learning experience. These kinds of technologies will support the creation of high-quality and adaptive educational resources, helping participants remain motivated and actively involved throughout the course. Specifically, AI tools will be employed to generate multimedia content tailored to diverse learners profiles, provide real-time feedback on assessment and learning activities through conversational agents capable of guiding participants through complex topics and answering frequently asked questions. In addition, reflective learning prompts, automatically generated by AI, will encourage students to connect course content with their own professional goals and real-word contexts. These strategies aim to create a more interactive and responsive learning environment, aligned with the pedagogical principles of engagement, relevance, and autonomy.

(c) The hybrid workshops: combining face to face sessions with online components. These workshops serve as applied learning spaces where participants engage in real-world tasks, including the technical issuance of credentials, integration into LMSs, and adaptation to institutional processes. In collaboration with international experts from UC3M and TU Graz, as well as local specialists, these workshops will address both global best practices and contextual challenges in Guatemala. Through case studies, demonstrations, and hands-on exercises, participants gain practical experience with digital credential systems, reinforcing their confidence and competence.

The implementation strategy also includes targeted outreach to employers, highlighting how digital credentials can enhance recruitment processes, improve talent matching, and contribute to a culture of lifelong learning.

4 Stakeholder Perspectives

The last section describes the training and implementation concept of how to introduce digital credentials in Guatemala. Though they become more and more integrated into formal and informal learning environments in some parts of the world, it is important to understand how various stakeholders perceive their value, utility, and limitations in the regional setting of Latin America. This section investigates distinct perspectives shaped by practical experience and expectations of students and experts that also need to be considered for the implementation of a digital credentialing ecosystem.

4.1 Students´ Perspective

Previous research has shown that students value opportunities to demonstrate their skills and achievements and to distinguish themselves from other candidates in competitive educational or employment contexts (Miller et al., 2017; Kiiskilä et al., 2023). Digital credentials, by making learning outcomes visible, verifiable, and portable, appear to align well with students' growing interest in flexible, skill-based recognition systems. For learners navigating increasingly complex educational and labor market environments, digital credentials offer a chance to document not only formal qualifications but also informal and non-formal competencies, thereby expanding their ability to signal expertise and readiness to employers.

To better understand students' acceptance of digital credentials in practice, we conducted a survey among students from Kinal, a vocational training institution that is a partner in the EcoCredGT project. An online questionnaire was administered to 113 male students that enrolled in five different vocational programs at Kinal. The students' ages ranged from 17 to 20 years with a mean of 17.9 years (SD = 0.57). The questions measured their perceptions of digital credentials, focusing on their intention to adopt them, the perceived utility for their careers, and any associated concerns.

The majority of students (81%) expressed a willingness to accept both paper-based and digital credentials, while only 5% preferred exclusively paper certificates and 13% would only opt for the digital credential. Students reported high confidence in their ability to effectively use new technologies and to handle digital credentials well in the future. Moreover, high ratings were given for the interest in receiving digital credentials to document their skills and achievements, to apply for jobs, and to support their career in general (see Table 1).

Factor	M	SD
Perceived ability to use new technologies effectively	4.38	0.70
Perceived ability to handle DCs well in the future	4.35	0.73
Interest in receiving DCs to document skills and achievements	4.47	0.78
Interest in DCs to use them for job applications	4.47	0.76
Perceived overall usefulness for career	4.53	0.70
Perceived cost (e.g. time, effort, resources)	4.35	1.00

Tab. 1: Mean rating scores (n=113). Note. For each factor the degree of agreement was measured on a Likert scale from 1 (not at all) to 5 (very).

Despite this overall positive reception, some students acknowledged potential challenges. The average perceived cost that needs to be invested to make digital credentials useful for oneself - defined in terms of time, effort, and resources - was moderate (see Table 1). A minority of students (12%) expressed specific concerns. On a personal level, these included confusion about the concept and functionality of digital credentials in general, or future disadvantages if people chose not to adopt them. At the institutional or systemic level, students raised issues related to data security, the recognition of previously earned certificates, and uncertainty about whether educational institutions or employers would value digital credentials. Additionally, concerns were voiced about the comparability and standardization of digital credentials across different settings, highlighting the need for clearer communication and institutional support to build trust in digital certification systems. Additionally, cost - both financial and in terms of effort - was perceived as a potential barrier by some respondents.

4.2 Experts´ Perspective

A panel discussion with four experts from different academic and professional backgrounds - STS, computer science and education - was conducted to gain broader insights into the potential and limitations of digital credentials. In detail, the involved people are affiliated with the Science, Technology and Society Unit at TU Graz (male), the Interdisciplinary Research Center for Technology, Work, and Culture (IFZ) in Graz (female), the Galileo University in Guatemala City (male) and the Universidad Carlos III de Madrid.

The panel highlighted the potential of digital credentials to expand access to education by recognizing smaller learning units and informal learning experiences. This modular

approach was described as an alternative to traditional, long-term educational programs and may facilitate participation among groups who are typically underserved by formal education systems. Digital credentials were seen as means to validate vocational and skill-based learning, particularly in contexts where such competencies often remain uncertified. The flexibility and portability of digital credentials were emphasized as key features that can support learner mobility and lifelong learning across institutional and national boundaries. Additionally, digital credentials were discussed as part of broader socio-technical transformations in education. Their implementation requires not only technological infrastructure but also alignment with institutional practices, cultural norms, and learner expectations. Overall, a great potential was seen for microcredentials in and for developing countries to support and train underprivileged groups.

Despite their potential, a number of critical challenges were raised. The importance of incorporating more diverse perspectives was emphasized as a key area for further attention. Furthermore, holders of digital credentials often wish to display their achievements publicly - such as on social media or professional networking platforms - to demonstrate their achievements. However, the visibility of such credentials may be influenced by the business models of these platforms, which often prioritize content from users who pay for premium features. This raises concerns about digital exclusion, as not all users can afford or access such visibility-enhancing options. Consequently, digital credential ecosystems must be designed to be open, fair, and resilient, avoiding over-reliance on a few commercial platforms that may amplify inequalities rather than reduce them. Other concerns included the recognition of digital credentials by institutions and employers, the comparability of credentials across systems and regions, and issues related to data privacy and digital security. Additionally, while digital credentials may reduce certain barriers to participation, the initial implementation of such systems could unintentionally exclude some groups, if they don't have access to the required digital technology. This tension between the goal of inclusivity and the realities of staged implementation was acknowledged as a challenge requiring ongoing reflection and adjustment.

Looking ahead, the experts agreed that the development and adoption of digital credentials must be approached as a dynamic and iterative process. Future efforts should prioritize broader inclusion by engaging underrepresented groups, addressing access and usability issues, and fostering trust through transparency and institutional support. This includes the incorporation of clear standards for credential recognition, user-friendly platforms, and supportive policies that encourage uptake among learners and employers. Moreover, the panelists emphasized that digital credentials should not be viewed solely as technical solutions but as components of a broader educational and social ecosystem. Their success will depend on continuous dialogue between educators, technologists, learners, and policymakers and on a shared commitment to ensuring that digital innovation supports equity, accessibility, and meaningful learning.

5 Discussion and Limitations

The last section provided an impact analysis from the perspectives of students and experts. The findings from the Kinal student survey indicate strong interest and acceptance of digital credentials, particularly with regard to their potential to support career development. Also the experts highlighted the potential of digital credentials for supporting learner mobility and lifelong learning across institutional and national boundaries.

However, both students and experts also raised some concerns. First, students and experts had some doubts regarding data privacy and digital security. In fact, as described in Section 2.1 digital credentials are designed in a way that they provide high security standards, independent if implemented centralised or decentralised. The training units created for students, educators, and administration and technical staff also include information about digital security to clarify issues on data protection.

Second, students and experts questioned the sustainability of digital credentials, as they were not sure if they are recognised by institutions and employers across systems and regions. Again, Section 2.1 presents various frameworks and standards related to digital credentials, which is also included in the training units for operators and administrative staff. This is an attempt to raise awareness for the importance of standards to ensure interoperability and sustainability of credentials. However, it is out of scope of our project to guarantee that obtained digital credentials are accepted always and everywhere.

Third, similar to the recognition issues, the experts also raised concerns that the visibility and accessibility of digital credentials on platforms depends on the platform where they are stored. Although our project does not have influence on digital credential platforms and their business models, there is still the opportunity to make students and also institutions aware of the importance of visibility and accessibility. Furthermore, digital credentials can also be stored on a decentralised block-chain, which provides equal and fair conditions for everybody, and removes dependency on platform providers and their business models.

Finally, the danger of unintentionally excluding learner groups due to a lack of access to the required digital technology in the transition phase when introducing digital credentials. This certainly requires careful planning of the initial implementation phase that monitors difficulties, obstacles, and fails when dealing with digital credentials. In our project, the Fundación Kinal is experienced in managing the access to learning content to diverse social groups, which also is also beneficial for organising fair access to digital credentials. Furthermore, in Latin America most people have at least access to a Smart Phone that can be used for receiving and managing digital credentials, if planned in advance.

6 Conclusion and Outlook

Digital credentials are increasingly recognized by both learners and experts as valuable instruments for the flexible and transparent recognition of skills and achievements. This article presents the initiative of the Erasmus+ project EcoCredGT that aims to implement digital credentials in Guatemala. This pilot project seeks to demonstrate how digital credentials can be established in Latin America. Therefore, a holistic training concept and course has been elaborated that addresses the whole development cycle including the creation of learning content, its deployment, and the issuance of digital credentials. Furthermore, the deployment of this course is presented consisting of a webinar, a Moodle course, and online workshops. The target audience is three-fold by addressing students, teachers, and administrative staff, which ensures that all stakeholders are made aware of a proper deployment of digital credentials. In an analysis, students and experts reacted very positively on this approach. However, some concerns raised - ranging from conceptual understanding and perceived costs to questions of recognition and institutional support - underscore that enthusiasm alone is insufficient for widespread adoption.

As a next step a digital credentialing system will be set up that complements the theoretical knowledge of the training course with practical experience of issuing and receiving digital credentials. Bridging the gap between the theoretical promise and practical implementation of digital credentials will require targeted and coordinated efforts across multiple levels. Effective communication, educational outreach, and policy development are essential to building awareness, trust, and usability. Ultimately, the success of digital credentialing systems will rely on inclusive and collaborative strategies that ensure these tools are not only technically robust, but also socially equitable and pedagogically meaningful.

Acknowledgement

This work has received funding from the European Union's Erasmus+ programme under grant agreement No. 101129122 (EcoCredGT). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency (EACEA). Neither the European Union nor EACEA can be held responsible for them.

References

100.000 Strong in the Americas (2024). *SENA (Servicio Nacional de Aprendizaje)*. Retrieved from https://www.100kstrongamericas.org/sena/

1EdTech (2024). *Open Badges Specification Conformance and Certification Guide*. Retrieved from https://www.imsglobal.org/spec/ob/v3p0/cert/

Angélica Ducci, M. (1997). New challenges to vocational training authorities: Lessons from the Latin American experience. *International Journal of Manpower*, 18(1/2), 160–184. https://doi.org/10.1108/01437729710169328

American Association of Collegiate Registrars and Admissions Officers (AACRAO) (2023). *Credential Confusion: A Call for Uniformity in Practice and Terminology*. Retrieved from https://www.aacrao.org/research-publications/aacrao-research/credentials-confusion-a-call-for-uniformity-in-practice-and-terminology

Amado-Salvatierra, H. R., Morales Chan, M., & Hernandez-Rizzardini, R. (2024). WIP: ECOCredGT Implementing Digital Credentials in Continuous Training for the Labour Market. In *2024 IEEE Frontiers in Education Conference (FIE)* (pp. 1-5). IEEE.

Brands, S. (2002): *A Technical Overview of Digital Credentials*. Retrieved from http://www.credentica.com/overview.pdf

Brown, M., Nic Giolla Mhichil, M., Beirne, E., & Mac Lochlainn, C. (2021). The Global Micro-credential Landscape: Charting a New Credential Ecology for Lifelong Learning. *Journal of Learning for Development*, *8*(2), 228-254. doi:10.56059/jl4d.v8i2.525

Cinterfor/ILO. (2001). Modernization in Vocational Education and Training in the Latin American and the Caribbean Region. Montevideo: ILO/Cinterfor.

Confederação Nacional da Indústria (n.d.). *Exporte seu produto com segurança e competitividade*. Retrieved from https://www.portaldaindustria.com.br/cni/canais/assuntos-internacionais/o-que-fazemos/solucoes/certificado-de-origem-digital/

Delgado Kloos, C. et al. (2025). How Challenges Become Opportunities: Micro-credentials and Artificial Intelligence. *IEEE EDUCON 2025 Conference,* London, UK, 22-25 April 2025, pp. 1-10. DOI: 10.1109/EDUCON62633.2025.11016509.

Erasmus+ Programme Guide. Glossary of terms - Vocational Education and Training. (n.d.). Erasmus+. https://erasmus-plus.ec.europa.eu/programme-guide/part-d/glossary-vet?

European Commission (2020). Final report: A European approach to micro-credentials. Output of the Microcredentials Higher Education Consultation Group. Retrieved from https://ec.europa.eu/education/sites/default/files/document-library-docs/european-approach-microcredentials-higher-education-consultation-group-output-final-report.pdf

European Commission (2023). *Europass: Improve your Career and Learning Opportunities*. Retrieved from https://europa.eu/europass/en/about-europass

European Commission (2024). *eIDAS Regulation*. Retrieved from https://digital-strategy.ec.europa.eu/en/policies/eidas-regulation

Eur-Lex (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council on the on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). *Official Journal of the European Union*. Retrieved from https://eur-lex.europa.eu/eli/reg/2016/679/oj

Foshay, W. R., & Hale, J. (2017). Application of principles of performance-based assessment to corporate certifications. *TechTrends*, *61*(1), 71-76. https://doi.org/10.1007/s11528-016-0125-5

Goldreich, O., Micali, S., & Wigderson, A. (1986). Proofs that yield nothing but their validity and a methodology of cryptographic protocol design. *27th Annual Symposium on Foundations of Computer Science*, 174-187. doi: 10.1109/SFCS.1986.47

Gräther, W., Kolvenbach, S., Ruland, R., Schütte, J., Torres, C., & Wendland, F. (2018). Blockchain for Education: *Lifelong Learning Passport* [Conference paper]. Proceedings of 1st ERCIM Blockchain Workshop 2018, Amsterdam, Netherlands. doi: 10.18420/blockchain2018_07

Grech, A., & Camilleri, A. F. (2017). Blockchain in education. *Joint Research Centre of the European Commission*. doi:10.2760/60649. Retrieved from https://publications.jrc.ec.europa.eu/repository/handle/JRC108255

Grech, A., Sood, I., & Ariño, L. (2021). Blockchain, self-sovereign identity and digital credentials: promise versus praxis in education. *Frontiers in Blockchain*, *4*, 616779.

Hernandez, R. et al. (2014) Promoting engagement in MOOCs through social collaboration Oxford UK: Proceedings of the 8th EDEN Research Workshop.

Hirsch, D. (2025). From global trends to national specificities in Vocational Education and Training: empirical and methodologic al contributions from a Latin-American case study. *Journal for Critical Education Policy Studies (JCEPS)*, *22*(3).

IMS Global. (2025). *Open Badges 3.0 Implementation Guide: Final Release Spec Version 3.0*. Retrieved from https://www.imsglobal.org/spec/ob/v3p0/impl/

Kato, S., V. Galán-Muros & T. Weko (2020). The emergence of alternative credentials. *OECD Education Working Papers*, *No. 216*. https://doi.org/10.1787/b741f39e-e

Katz, J., & Lindell, Y. (2007). *Introduction to Modern Cryptography: Principles and Protocols*. Chapman and Hall/CRC Press. https://doi.org/10.1201/9781420010756

Keevy, J. and Chakroun, B. (2015). *Level-Setting and Recognition of Learning Outcomes: The Use of Level Descriptors in the Twenty-First Century*. Retrieved from https://doi.org/10.54675/GKWN6283

Kemcha, R., Alario-Hoyos, C., & Delgado-Kloos, C. (2024). Exploring Recognition in Digital Education through Open Badges and the European Learning Model. In *2024 IEEE Digital Education and MOOCS Conference (DEMOcon)* (pp. 1-6). IEEE.

Lemoie, K. (2024, October 7). Explaining Verifiable Credentials and Open Badges 3.0: Part 1: The Trust Model of Open Badges. *Digital Credential Consortium*. Retrieved from https://blog.dcconsortium.org/explaining-verifiable-credentials-and-open-badges-3-0-5bf2f482b383

Lemoie, K. (2025, January 15). Explaining Verifiable Credentials and Open Badges 3.0: Part 2: Issuing Badges. *Digital Credential Consortium*. Retrieved from https://blog.dcconsortium.org/explaining-verifiable-credentials-and-open-badges-3-0-34ae898b98b2

Liu, R. et al. (2025). *Improving AI in CS50: Leveraging Human Feedback for Better Learning.* SIGCSE TS 2025, Pittsburgh, PA, USA, 26 Feb-1 Mar 2025. https://cs.harvard.edu/malan/publications/fp0627-liu.pdf

Maclean, R., & Lai, A. (2011). Editorial: The future of technical and vocational education and training: Global challenges and possibilities. *International Journal of Training Research*, 9(1-2), 2–15. https://doi.org/10.5172/ijtr.9.1-2.2

Miller, S. (2024). What Is Vocational Training - Education, Program and Schools. https://www.vocationaltraininghq.com/what-is-vocational-training/

Morales Ramos, S., Carneiro, F., Castillo, M., Cattivelli, M., & Méndez, G. (2022). Juventudes vulnerables, competencias digitales y formación profesional en América Latina. (OIT/Cinterfor). 72p. https://www.ilo.org/es/publications/juventudes-vulnerables-competencias-digitales-y-formacion-profesional-en-0

Mühle, A., Assaf, K., Köhler, D., & Meinel, C. (2023). *Requirements of a Digital Education Credential System* [Conference paper]. IEEE Global Engineering Education Conference (EDUCON), Kuwait. doi: 10.1109/EDUCON54358.2023.10125183

Narayanan, A., Bonneau, J., Felten, E., Miller, A., & Goldfeder, S. (2016). *Bitcoin and Cryptocurrency Technologies: A Comprehensive Introduction*. Princeton University Press. https://doi.org/10.1016/S1353-4858(16)30074-5

Pontificia Universidad Católica del Perú (2024). *Oferta académica*. Retrieved from https://educacioncontinua.pucp.edu.pe/oferta-academica/

Porto, S. & Presant, D. (2023). *The IDB Digital Credential Framework: Principles and Guidelines for Creating and Issuing Credentials*. Retrieved from https://publications.iadb.org/en/publications/english/viewer/The-IDB-Digital-Credential-Framework-Principles-and-Guidelines-for-Creating-and-Issuing-Credentials.pdf

Prada, M. F., & Rucci, G. (2023). Skills for Work in Latin America and the Caribbean: Unlocking Talent for a Sustainable and Equitable Future. Publications IADB.

Preukschat, A., & Reed, D. (2021). *Self-Sovereign Identity: Decentralized digital identity and verifiable credentials*. Manning Publications. ISBN: 9781617296598

Quigley, J. (2023). *What Are Digital Credentials and How Are They Used?* Accredible. Retrieved from https://www.accredible.com/blog/what-are-digital-credentials

Salazar-Xirinachs, J. M., & Vargas Zúñiga, F. (2017). The future of vocational training in Latin America and the Caribbean: overview and strengthening guidelines. *Montevideo: OIT/Cinterfor*.

Sedlmeir, J., Smethurst, R. & Rieger, A. (2021). Digital Identities and Verifiable Credentials. *Business & Information Systems Engineering*, *63*, 603–613. https://doi.org/10.1007/s12599-021-00722-y

Suescún Barón, C. A., Hernández Pérez, S. S., Giraldo Pedroza, J. S., & Tellez Pérez, J. D. (2024). Vocational education and training in Latin America. *RBEST Revista Brasileira De Economia Social E Do Trabalh*o, 6, e024013. https://doi.org/10.20396/rbest.v6i00.19973

Tecnológico de Monterrey (2023). *Tec de Monterrey launches digital credentials to certify skills*. Retrieved from https://conecta.tec.mx/en/news/national/education/tec-de-monterrey-launches-digital-credentials-certify-skills

Todd, R., & Dunbar, M. (2018). Taking a whole of government approach to skills development. UNESCO; *International Labour Organization.* https://doi.org/10.54675/SGOS3486

UNESCO (2018). *Digital Credentialing Implications for the recognition of learning across borders*. Retrieved from https://oer4nosp.col.org/id/eprint/16/1/264428eng.pdf

UNESCO (2022).*Towards A Common Definition Of Micro-credentials*. Retrieved from https://unesdoc.unesco.org/ark:/48223/pf0000381668#:~:text=Micro-credentials%20are%20often%20promoted%20as%20an%20efficient%20way,standards%20and%20is%20awarded%20by%20a%20trusted%20provider

Universidad de Chile (n.d.). *Information about the revalidation and recognition processes at the University of Chile*. Retrieved from https://uchile.cl/english-version/international-relations/information-about-the-revalidation-and-recognition-processes

World Wide Web Consortium (2022a). *Decentralized Identifiers (DIDs) v1.0*. Retrieved from https://www.w3.org/TR/did-core/

World Wide Web Consortium (2022b). *Verifiable Credentials Data Model 1.0*. Retrieved from https://www.w3.org/TR/vc-data-model

World Wide Web Consortium. (2024). *Verifiable Credentials Data Model v2.0*. Retrieved from https://www.w3.org/TR/vc-data-model-2.0/

Yli-Huumo, J., Ko, D., Choi, S., Park, S., & Smolander, K. (2016). Where is current research on blockchain technology? A systematic review. *PloS One*, *11*(10), e0163477. https://doi.org/10.1371/journal.pone.0163477

Enabling Dilemma of AI for Disabled Individuals

Narges Naraghi[1,2], Linda Nierling[1], Matthias Wölfel[2,3]

[1] Karlsruhe Institute of Technology, Germany

[2] Karlsruhe University of Applied Sciences, Germany

[3] University of Hohenheim, Germany

DOI 10.3217/978-3-99161-062-5-018, CC BY 4.0

Abstract. The fairness of Artificial Intelligence (AI) for individuals with disabilities is a complex and contested issue, as AI holds both inclusive and exclusive potential. On the one hand, AI can empower disabled individuals by mitigating barriers; on the other hand, it may perpetuate discrimination against marginalized groups, including those with disabilities. Intersectionality further differentiates this picture by highlighting how multiple forms of discrimination intensify these challenges. Leaning on this argument, this paper addresses the following question: How do intersectional forms of discrimination interfere with the enabling power of AI for disabled individuals? We argue that autonomy, the capacity to decide, plan, and act toward personal goals, provides a fitting analytical lens, as it encompasses crucial dimensions like agency and accessibility. Using a qualitative analysis of 48 online documents publicly available at websites that address inclusive AI for disability, we identify two key insights. First, intersectional discrimination does not merely obscure AI's enabling potential; it can actively reverse it, undermining the autonomy of disabled individuals. Second, bringing the broader society into the analysis, the control of disabled people over their lives, as compared to the society that they live in, may shrink, regardless of their autonomy in their personal lives. This debate formulates AI's enabling dilemma: while promising empowerment, AI may deepen disparities due to intersectionality and the accelerating enablement of the general population. Fairness of AI, therefore, must be assessed not only through the lens of disability but also in the context of broader societal structures and inequalities.

1 Introduction

Artificial Intelligence (AI) is increasingly becoming a pivotal force in evolving technologies that support disabled individuals[1], emerging through two parallel but converging trends: pursuing accessibility for mainstream consumer AI technologies and shifting dedicated assistive technologies into AI-based solutions (Braun et al., 2020). On the one side, consumer technologies, especially those developed for mass markets, progressively try to incorporate features enhancing accessibility. On the other side of the spectrum lie purpose-built assistive technologies explicitly designed to support disabled individuals. However, promises of solutions suggested by both sides of the spectrum are far from fully realized due to technical challenges and socio-technical contexts. Nonetheless, while necessary, ongoing debates about algorithmic bias and fair access are not sufficient to reach fairness, understood as a social good (Lillywhite and Wolbring, 2023) and justice (Hertweck et al., 2024). Instead, we would argue that fairness must be understood in broader terms, extending beyond algorithmic performance to encompass real-world social conditions and lived experiences. Accordingly, this paper takes a step beyond technical aspects of fairness (such as Pagano et al., 2022; Mehrabi et al., 2019) to critically explore whether AI systems can truly function as enabling technologies for disabled individuals in everyday life.

Central to this inquiry is the concept of intersectionality: a framework that reveals how overlapping forms of discrimination, such as those based on disability and ethnicity, can interact to shape unique and often intensified experiences of marginalization (Wolbring and Nasir, 2024). In this context, we encounter two opposing dynamics: on the one hand, AI holds potential as an enabling tool for disabled individuals; on the other hand, structural and intersectional forms of discrimination may limit or even negate this potential. So, this study shifts focus from technical remedies to a qualitative, socio-technical analysis of AI's enabling role. We aim to inquire how disabled individuals experience AI in their daily lives, particularly within settings marked by intersecting axes of inequality. Therefore, the main argument would be whether AI can serve as a substantial enabler or risks reproducing existing forms of exclusion in new, technologically mediated ways. To be more specific, we would investigate how the experience of using AI would be affected by additional forms of discrimination, such as for ethnical minority social groups, in addition to disability.

To meaningfully assess the enabling power of AI for disabled individuals, it is both conceptually and ethically justifiable to focus on its potential to enhance autonomy. This focus could be rooted in the terminological nexus between ‘ableism’ and ‘enabling.’

[1] We acknowledge two politically correct approaches to address people with disability (and disabled individuals). We have used both depending on articulation of the sentence, given priority to the phrase disabled individuals.

Ableism, as defined in disability studies, refers to the systemic discrimination and social prejudice directed toward individuals with disabilities, manifesting in practices that marginalize disabled voices, reduce individuals to their impairments, and uphold normative assumptions about ability (Hofmann et al., 2020; Shew, 2020). In contrast, the notion of enabling represents a reorientation: it emphasizes empowering disabled people (either disabled due to impairment or by society[2]) and supporting independent living (Moyà-Köhler and Domènech, 2022) or, as Wolbring (2024) depicts, turning expectations around ability and ableism into opportunities for empowering individuals and reshaping social structures.

Within this conceptual shift, autonomy emerges as a particularly salient value. Nonetheless, autonomy also encapsulates itself as a spectrum of interrelated dimensions in the neighbouring disciplines. To begin with, from a philosophical perspective, autonomy is a key value in human identity, as emphasized by Kant (Chiodo, 2022), and it features prominently in technology design methodologies such as Value-Sensitive Design (VSD), where it is treated as a key ethical value that developers should aim to preserve and enhance (Friedman and Hendry, 2019). Meanwhile, technology could also aid in obtaining and maintaining autonomy, where autonomous self-realization and human agency, among others, are listed as opportunities brought by AI to society (Floridi et al., 2018). Thus, framing the enabling potential of AI in terms of its capacity to promote autonomy is not only consistent with the aims of disability advocacy but also aligns with broader ethical and design principles in the development of emerging technologies. It allows us to examine if AI can serve as a medium for empowerment rather than as a new vector of dependency or exclusion. Accordingly, this study hypothesizes that AI has some potential to enable disabled individuals by raising their autonomy, but this enabling power should be further differentiated, taking intersectionality into account. Therefore, this study aims to qualitatively analyze the interrelation between employing AI, the autonomy of disabled individuals, and intersectionality to answer the question of ‘how intersectional forms of discrimination affect the autonomy of disabled individuals while using AI’.

Our study, thereby, turns abstract technological potential into tangible social progress by bringing intersectional discrimination into the analysis of the enabling power of AI. In this framing, disability is not treated in isolation, nor is technology viewed as a neutral or universally empowering tool. Rather, we foreground the everyday life of disabled individuals, where intersectional forms of discrimination are not peripheral but integral, and investigate how AI functions as an enabling variable within this complex terrain. Thus, the core aim of this research is to qualitatively analyze the dynamics between AI, autonomy, and intersectional discrimination to assess whether and how AI can support

[2] This dual definition of disability roots in the two predominant models of disability: the medical model, which associates disability with physical impairment, and the social model that finds society and the environment as disablers (Mitra, 2006).

more autonomous lives for disabled individuals, considering multiple forms of discrimination.

To answer this question, Chapter 2 focuses on the theoretical background of the research, which is followed by a methodology section, Chapter 3, where the path of selecting the research method and empirical work is reported. Accordingly, the results and the discussion of the research are reported respectively in Chapters 4 and 5. The paper concludes with the research contribution and further research in Chapter 6.

2 Theoretical Framework of the Research

This chapter elaborates on two key theoretical standpoints of this study: how autonomy could be understood from the literature, and how layers of discrimination could affect the enabling potentials of AI.

2.1 Autonomy and Disability

As a core value in contemporary ethics, autonomy is both a complex and context-sensitive term. At its foundation, autonomy is understood as an individual's capacity to act based on their own beliefs, motivations, and values, free from coercion, manipulation, or deceptive influence (Prunkl, 2024). This notion includes both the authority and power to live one's life (Prunkl, 2024). Within the field of value-sensitive design, autonomy is further conceptualized as the ability of individuals to decide, plan, and act in ways that support their self-defined goals (Friedman and Hendry, 2019). Importantly, these perspectives converge on the view that both internal authenticity (actions reflecting one's true self) and external agency (the actual capacity to act meaningfully in one's environment) are essential in defining autonomy (Prunkl, 2024). Alternatively, Laitinen and Sahlgren (2021) address these two aspects as human autonomy and functional autonomy, where the former incorporates the latter with an adequate degree of control, and the latter is responsible for operating independently.

However, the lived experience of autonomy diverges significantly between abled-bodied and disabled individuals. To dive into this difference, it would be helpful to first elaborate on Chiodo's (2022) perspective on distinct human autonomy from technological automation; while the former is in the hands of the person, the latter is beyond their control, or, as Chiodo articulates, off-hand or outsourced to the machine. Speaking of disabled individuals, the second pillar could be considered as their subordinate hand[3] (Moyà-Köhler and Domènech, 2022) or an extension of their body, illustrating how

[3] This expression uses ‘hand” as a representative metaphor for agency; and how technology for disabled individuals could be considered as their secondary ‘hand” in order to compensate on the limitations the disability may cause.

autonomy for these individuals is often mediated by the technical form and function of the tools at their disposal. This position of technology fits well with some philosophical definitions of technology. On top of them, McLuhan's (1994) definition of technology as an extension of man, Latour's (2005) Actor-Network Theory when he realizes the agency of non-human actors (as actants), not to mention Haraway's (2016) Cyber Manifest where the distinction between human and machine is being questioned, support this proposition that technology for a disabled individual is more than a tool. Instead, it becomes part of their body and gives them the freedom to carry their lives (Mazera *et al*., 2024) and facilitates greater autonomy (Moyà-Köhler and Domènech, 2022).

This brief review of autonomy, how it is perceived differently for non-disabled versus disabled individuals, and how technology and automation manifest for them, is an essential theoretical backbone for this study. This basis provides us with an understanding of the autonomy of persons with disability to assess the enabling power of AI through our empirical data.

2.2 Intersectional Discrimination in the Disability Realm

Our main argument in this section is to depict how the intersections of discrimination affect the realization of the potential and promises of digital technologies, such as AI. In this regard, as shown in Fig. 1., we analyze different layers of discrimination and exclusion from AI technologies. This pyramid shows how each layer of discrimination can diminish the enabling power of AI. As mentioned above (while articulating the research problem in the introduction section), enabling power could be understood as reversing the disabling attributes of the physical impairment and society, or, in a broader context, the autonomy of individuals.

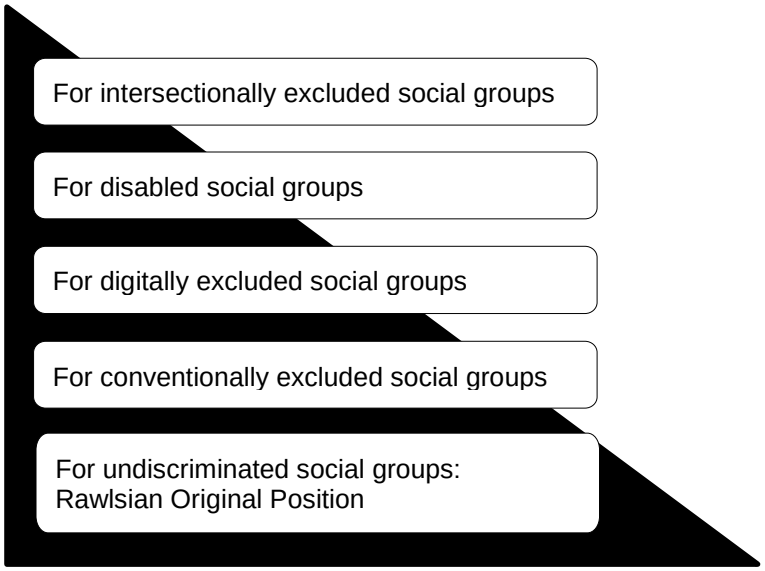


Fig. 1.: How different layers of discrimination diminish the enabling power of AI- Own presentation, inspired by (Park and Humphry, 2019).

John Rawls' (1971) theory of *the Original Position* or *the Veil of Ignorance* could perfectly define a purely inclusive world where decision-makers (here, technology developers, technology policy-makers, investors, etc.) ignore their gender, class, nationality, and any other identifying attributes they might carry. This positioning fits well with the definition of fairness in the decision-making context, which is the '*absence of any prejudice or favoritism toward an individual or group based on their inherent or acquired characteristics'* (Mehrabi et al., 2019, p. 1). According to Rawls, to diminish the bias, one should return to one's original position before birth. Something that is neither feasible nor plausible in the real world, where Southern citizens, marginalized ethnicities, the lower economic classes, the older adults, women, non-binary persons, and, in our case, people with disabilities, are facing discrimination in many spheres of their lives, including utilizing new technologies. Nevertheless, these conventional exclusionary topics are not the only discriminating attributes affecting the utilization of digital technologies. Digital divide or digital exclusion, limited or lacking access to the Internet, smart devices, and other infrastructures, and a lack of the authority to use digital tools and features (Park and Humphry, 2019) are all discriminatory aspects that may be imposed on someone, in addition to the conventional forms of discrimination. In addition to these layers, AI systems may impose particular demands on users that persons with disabilities, as users, might not be able to meet. For example, speech-based interfaces often require clear articulation and the ability to formulate precise commands or intents. These requirements can pose significant challenges for individuals with speech impairments, cognitive limitations, or language barriers. As a result, such interfaces risk excluding or disadvantaging certain user groups unless alternative interaction modalities or inclusive design principles are implemented.

It is worth mentioning that these layers are defined and divided in the related literature in different arrangements; for instance, some researchers consider the digital divide as an extension of general exclusion (Nierling and Maia, 2020). Others merge the exclusion of people with disabilities with the digital divide layer (Braun et al., 2020).

Regardless of how these layers are addressed, there is a hidden layer that is manifested through the intersection of two or more discriminations. As developed by some scholars, intersectionality elaborates on the idea that different attributes of identity shape the form and change the dynamics of oppression (Wolbring and Nasir, 2024). Intersection of multiple forms of discrimination, hence, further questions the fairness of AI for disabled individuals (Lythreatis et al., 2022; Mitra, 2006; Tsatsou, 2020). In other words, being classically discriminated against (such as exclusion caused by gender, religion, race, etc.), being excluded from digital technologies (such as data exclusion, algorithmic biases, etc.), and lacking access due to disability (such as inaccessible UI designs, etc.) are not independent of each other. Instead, the intersection of discriminations counteracts each other to shape a new dynamic of discrimination. To be more specific, the experience of discrimination for any given disabled person could be different from

others, given that they might face multiple forms of discrimination. Accordingly, people facing the intersection of various discriminations might experience AI and utilize its potential differently from one another.

3 Methodology, Method, Empirical Work

This study employs a Critical Discourse Analysis (CDA) approach to examine the impact of AI technologies on the autonomy of individuals with disabilities, taking into account intersectional discrimination. CDA studies power imbalance, dominance, and discrimination through the use of language (Mengibar, 2015), the interpretation of mutually linked texts and other sources (Bondarouk and Ruël, 2004), and uncovering hegemony beneath marginalized individuals and ideas (Wall et al., 2015). These attributes enable us to trace the dynamics of power, exclusion, and representation within a discourse, making it especially suitable for studies that involve forms of discrimination (Noble, 2020), in our case, ableism. In order to analyse the relevant public discourse, we conducted a qualitative analysis covering online documents published by actors pertinent to the field. For our study, we analysed text documents that were all available online (e.g., news, articles, or blog posts at publicly available websites), covering a broad range of actors, ranging from public institutions to media channels to individual authors.

Accordingly, the empirical data for this study were selected via the DuckDuckGo search engine to make sure that the search is unbiased and not influenced by user-specific tracking or personalized algorithms. Aligned with our research question and methodology, we were looking for sources of intersectional forms of discrimination and how they affect the autonomy of disabled individuals in the public sphere. Accordingly, we needed to make sure that our sample is broad enough to provide a basis for critically analyzing the text and context, while also being relevant enough to our inquiry. Thus, we resolved to assemble a discourse that includes three key aspects: the (1) inclusion of (2) AI for (3) disability, which we used as our initial keywords for searching. The targeted search strategy, then, included the presence of the three keywords as mentioned above, their synonyms, or their inherent inclusion within the title of the entries. Here are the alterations of each keyword that helped us decide about the entries:

- Inclusion and its cognates (inclusive, inclusivity); fair and its cognate (fairness), equal opportunities, justice; representation, and not leave people behind.
- AI, generative AI; algorithms, algorithmic tools, language models; GPT-4, ChatGPT, OpenAI.
- Disabled people and different wordings for it; ableism, technoableism; and specific disabilities such as visually impaired, low-vision, etc.

The collection and analysis of the data were conducted from May to October 2024. The number of entries, when qualitative data saturation was reached, was 48, with different

scopes that they cover, publishing dates, and publishers, all of which are described below.

The piled-up sample covers a variety of AI technologies (language models; matchmaking algorithms; facial, voice, and motion recognition systems; robotics; etc.), provides narrations across various settings (everyday life, education, work, among others), and reflects a broad spectrum of disabilities (hearing, vision, mental, and physical impairments). The entries, those with a specified publishing date, range from 2019 to 2024, along with seven entries with no available date for publishing. Above that, to describe the context for the entries, which is expanded and inclusive, they can be categorized in five groups, based on the type of the publishing organization: accessibility and disability advocacy which represents publishers whose mission is addressing disability and accessibility (15 entries); blog, NGO, and independent authors for those who have no ties to public and private organizations (6 entries); business, technology, and market analysis for publishers with technical and market oriented views (7 entries); news and media outlet for publishers such as news agencies and analytical content providers (10 entries); and public institution and international organization for those who have ties to public institutes (10 entries). This information, along with the authors and publishing organizations, as well as access links, is all addressed in Annex 1.

The interpretation process was conducted by a single coder using Software support[4]. The coding process was carried out in two main stages, repeated after each round of data collection: applying codes to the text based on the key concepts of the research question (autonomy and intersectionality) and identifying emergent subcodes for each of these concepts. Additionally, since context plays a critical role in interpretation within the framework of CDA (Mengibar, 2015), we simultaneously applied an open coding strategy to contextual variables. Through this process, two principal contextual codes (emotional tone and attitudinal stance of the text) and some subcodes for each emerged, complementing the primary coding scheme. These codes and their subcodes are all addressed in Annex 2. The results of the research, accordingly, were derived through cross-analyzing these codes and concepts, seeking to identify how intersectional factors shape the autonomy and agency of disabled people.

4 Results

As mentioned above, while interpreting the qualitative data, two key dynamics became central to understanding the autonomy of disabled individuals in the context of AI. The first, intersectional discrimination, was a focus from the outset of the study, given its well-articulated impact on access, inclusion, and agency. However, a second emergent

[4] MaxQDA Analytics Pro. 2020

dimension also surfaced during the analysis: the relative autonomy of disabled individuals within the broader social environment. This additional factor underscores how autonomy is affected not only by technical offers of AI but also by one's embedded position in socio-technical contexts. This finding resonates with the social model of disability, which attributes disabling barriers primarily to society and social institutions rather than to individual impairments (Lawson and Beckett, 2020). These two dimensions offer a more layered and context-sensitive understanding of autonomy and of analyzing the promises of AI in general.

4.1 Addressing Autonomy Across the Intersectionality

Our analysis confirms that AI holds considerable enabling potential for people with disabilities, particularly in enhancing autonomy by supporting authentic decision-making and expanding agency. However, more critical interpretations of our data reveal that this potential is far from universally achievable. Accordingly, our empirical data includes some quotes to bring up the intersectional discriminations inherent in AI promises for people with disabilities: '*In some countries, immigrants tend to avoid medical examinations and tests for fear of being deported or facing unacceptable medical costs (46, public institution and international organization, 2023).'* On the contrary, '*Particular social groups (e.g., Caucasian families in the US) are more likely to report concerns related to the child's autism due to better medical access (46, public institution and international organization, 2023).*' Here, immigration background, ethnicity, and financial status, three different forms of discrimination, intersect with each other, affecting the seeking of medical solutions.

Exclusion from using technology, in our case AI, further compounds this problem: '*People with disabilities are one of the most marginalized groups in the effects of technology (8, accessibility and disability advocacy, NA).*' In other words, even when technologies exist and are accessible, they are not equally usable for all: '*mere existence of a[n AI] technology is not the same thing as people with disabilities having easy, affordable access to these things to actually use (21, blog, NGO, and independent, 2024).*' Technology is, as discussed above, the subordinate or secondary hand for disabled individuals, and access to it is not taken for granted. So, these observations feed this interpretation that the notion of AI as an 'enabler' is compromised when additional layers of discrimination, economic access, and digital literacy are factored into the equation. The experience of AI, accordingly, for a disabled individual who is also an ethnic minority, is a new form of discrimination shaped out of the intersection of two separate forms.

Another group of quotes is more implicit: by elaborating on the potentials and promises of AI for disabled individuals, they presume its availability and affordability. Take this quote, '*Social robotics for emotional training for pupils with autism [...] is a wearable that helps neurodiverse individuals with social-emotional learning (47, public institution and international organization, 2023).*' as an example. Interpreting through the context, as

CDA suggests, this quote encompasses multiple presumptions to conclude inclusivity of AI for disabled individuals; among them: pupils with autism can have access to social robotics; all pupils diagnosed in various spectrum of autism, with different language capabilities, mother tongues, and accents, can communicate with the robot; neurodiverse individuals have unified socio-emotional norms and subcultures. A similar quote states that '*AI powered robots and other tools […allow…] people with disabilities to live independently (43, public institution and international organization, 2022).*' This pattern has repeatedly occurred in our empirical data: many AI-based tools and features implicitly assume that users already possess certain forms of social and structural privilege, such as legal stability, economic means, and digital literacy.

Analysing quotes such as '*You can now find AI-powered braille tutor apps on the internet (25, business, technology, and market analysis, 2023)*' or '*The most common and affordable form of AI is using smart home technology (6, accessibility and disability advocacy, NA)*' implies a particular sentiment in which AI is there, and the disabled individual needs to utilize it as an enabler. At the same time, this baseline of technological access may not exist for all disabled individuals due to multiple and overlapping forms of marginalization.

The last but not least quote for this part is the one that offers a broad set of social attributes as disabling ones, implicitly suggesting that intersectionality is a disabler by itself, no matter how enabler AI is: '*Capitalism, racism, transphobia, patriarchy, colonialism, homophobia — all disabling (18, blog, NGO, and independent, 2024).*'

In summary, while AI may offer tools that could enhance the autonomy of disabled individuals, our findings show that an insufficient understanding of intersectional discrimination often undermines this promise. These layered exclusions limit AI's real-world availability and functionality, regardless of how accessible it might be. As a result, AI may not be as enabling in practice as it is often assumed or claimed to be in theory. This heightens the risk of disabled individuals becoming even further disabled in an AI-driven society, which is a key question in the following section of our results.

4.2 Addressing Relative Autonomy in the Society

A second aspect of the findings centers on the autonomy of disabled individuals as members of a society in which AI is an inseparable part. What this viewpoint suggests is that the perception of autonomy can be analyzed in a broader social setting. Regardless of the accessibility, availability, and affordability of AI for disabled individuals, AI-based solutions might still have a double effect on their autonomy because of the ableist mindset of society. In other words, for a person who faces no forms of discrimination except for their disability, fulfilling the promises provided by AI seems tricky: '*Police and autonomous security systems and military AI may falsely recognize assistive devices as a weapon or dangerous objects or misidentify facial or speech patterns (42, public*

institution and international organization, 2023).' This quote suggests that the ableist society could cancel out the autonomy offered by AI for disabled individuals.

Involving the person's social life (as opposed to their private life) in the perception of autonomy for disabled individuals also brings up the fact that disabled individuals are not the only social group that incorporates AI, and the whole of society also uses it. While non-disabled individuals are in the absolute majority in society, use tools and solutions of AI as decision-making aids, outsource their agency to the AI, and, in general, use AI as enablers, the expectations of performance for the entire society, including disabled and non-disabled social groups, escalate. This leaves those who do not have proper access to AI, due to merely disability or the intersection of disability and other forms of discrimination, behind: '*Educators already excessively discipline and punish […] disabled students, and stricter policing will exacerbate these disparities (17, blog, NGO, and independent, 2023).*'

Nonetheless, in the best-case scenario, when disabled individuals can utilize AI as expected, enabling promises of AI for disabled individuals might not be as welcome as anticipated by the entire society: '*ChatGPT threatens to disrupt able-bodied privilege (17, blog, NGO, and independent, 2023).*'

These insights point to broader societal shifts: while incorporating AI may suggest relatively higher living standards, it can simultaneously remove agency from disabled individuals by marking them as 'different.' Additionally, AI might be seen as threatening the privilege of non-disabled individuals, leading to a reconfiguration of autonomy and responsibility in a way that favors those who are already advantaged. Therefore, the enabling power of AI for disabled individuals must be compared to the dynamics of the entire society and not only within the narrow frame of disability-focused solutions, which could be either enabling or disabling.

5 Discussion

As mentioned above, despite the accessibility, availability, and affordability of AI for disabled individuals, AI-based solutions may still have a double-edged impact on their autonomy due to the prevailing ableist mindset in society. While AI holds considerable potential to enable disabled individuals, this potential remains unrealized mainly due to the persistent influence of intersectional discrimination. Rather than functioning uniformly as a tool for empowerment, AI can, in practice, reproduce or even intensify existing social inequities. Taking the nominal promises of AI to enable disabled individuals by raising their autonomy as the main argument of our study, two counter-arguments, as mentioned below and demonstrated in Fig. 2., raise serious doubts about the realizability of those potentials.

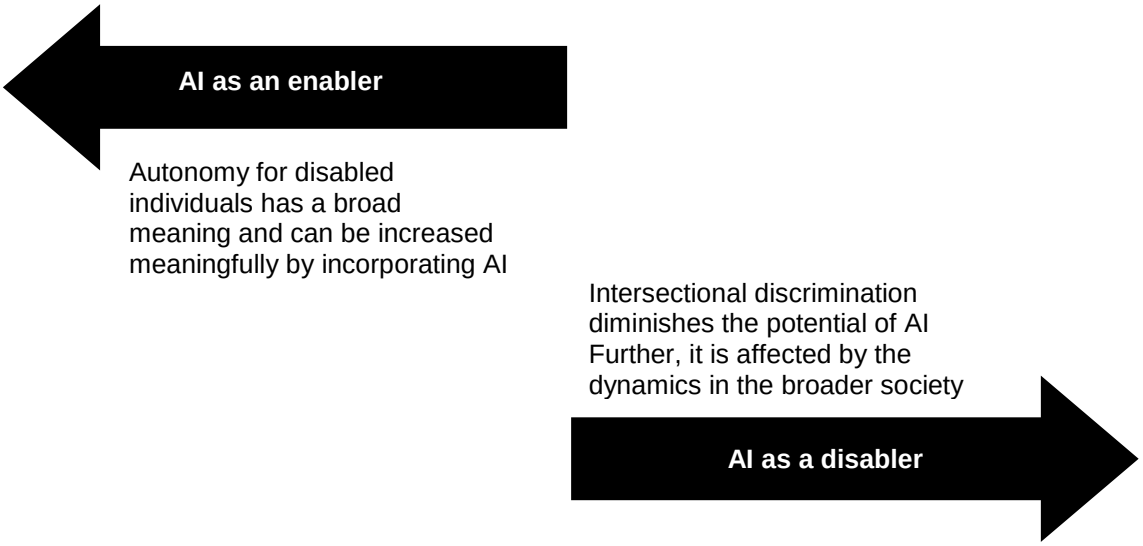


Fig. 2.: The enabling dilemma of AI- Own presentation

One counter-argument to the enabling narrative is that intersectionality itself may diminish the autonomy of disabled individuals when interacting with AI. As overlapping systems of oppression (such as racism, classism, or sexism) interfere with the equitable access and agency of disabled individuals, the enabling potentials of AI seem far-fetched. A second counter-argument, questioning the enabling promises of AI, emphasizes the importance of examining autonomy not in isolation but in relative terms. How AI restructures autonomy must be considered across the community, including disabled and non-disabled individuals. Therefore, a broader question arises: how, and if at all, can AI enhance the autonomy of disabled individuals compared to the society in which they live? Thus, any assessment of its enabling power ties to these relational dynamics, putting another layer of doubt on the enabling power of AI. These findings frame a dilemma: AI for disabled persons can act either as an enabler or a (further) disabler, depending on how incorporating it intersects with other forms of discrimination and in the broader social context. This dilemma is depicted in Fig. 2.

One practical contribution of this study is the emphasis on addressing not only disability-related bias in AI but also its intersection with other forms of discrimination that disabled users might face. Inclusion in AI development requires more than integrating disability perspectives into datasets or correcting for algorithmic exclusion. As Whittaker *et al*. (2019) argue, intersectionality fundamentally reshapes the operation of exclusion. We suggest that fairness in AI must be pursued through *intersectional debiasing*. By intersectional debiasing, we refer to the accounting for overlapping and mutually reinforcing effects either during model training or in post-processing. Effective intersectional debiasing must therefore recognize different layered identities (specifically those causing discrimination) and their complex interactions. However, like many inclusive AI efforts, intersectional debiasing faces trade-offs between inclusiveness and model performance. Data for multiply marginalized groups is often scarce, and its sparsity can hinder effective integration into training datasets.

The second core contribution of this study is suggesting a shift from evaluating AI's nominal potential to assessing the enabling capacity of AI in the everyday social settings in which disabled individuals live. We tend to call this *relative autonomy*, which, as Mazera *et al*. (2024) discuss, appears in everyday life in the face of external barriers that limit the actions of some people with disabilities. Timpe (2019) likewise reminds us that agency is not an isolated function of the individual but depends on the ecology and the environment. From a more prescriptive viewpoint, developing and using AI-driven consumer electronics and assistive technologies must be understood in context, as their development and use are influenced by local social, institutional, and cultural factors (Nierling and Maia, 2020). Similarly, as argued by Shams, Zowghi, and colleagues (2023), tackling bias and unfairness requires a holistic approach that recognizes the cultural dynamics and normative assumptions embedded within AI systems. Our contribution, then, ultimately aligns with shifting from D&I (Diversity and Inclusion) in AI to AI for D&I.

In conclusion, the enabling power of AI for disabled individuals cannot be assessed in isolation, separate from the community. It must be evaluated in light of social inequality, intersectional discrimination, and the contextual factors that shape autonomy and agency. Only through such a layered and situated approach can we begin to understand whether AI truly enables or disables the individuals it seeks to serve.

6 Conclusion and Further Research

This study shows that AI matters more for disabled individuals than is often presumed. Autonomy has a different operationalized connotation for disabled individuals, making AI and other promising technologies a potential leap in their quality of life. A non-disabled person might look at automation, giving out autonomy to the machine, as a trade-off that reduces their responsibilities (Chiodo, 2022). Meanwhile, technology for a disabled person is the freedom to carry out their lives (Mazera et al., 2024) and a promise of facilitating greater autonomy (Moyà-Köhler and Domènech, 2022). This distinction only applies to disability, as compared to other classic underrepresented social groups, making studying their autonomy a broad potential for further contribution.

Another theoretical standpoint of our study is to go beyond algorithmic fairness. To assess the fairness of AI for people with disabilities or any other underrepresented group, it is not sufficient to focus solely on the technical promises of AI or its specific applications for that group. Instead, the accurate assessment must be analyzed within a broader context to reach a comprehensive and meaningful evaluation of this socio-technical phenomenon. While this study is exploratory, it feeds further research that moves beyond assessing AI as a technical phenomenon and incorporates social dynamics as essential components in evaluating AI's enabling potential.

Based on these theoretical standpoints and through a qualitative analysis of empirical data, we identified and examined two key social forces that cast doubt on the promises of AI to enhance the autonomy of people with disabilities: intersectional discrimination and relative autonomy within broader society.

The latter expresses the need to clarify what some might expect from AI as an enabler. Is it expected of AI to maintain the status quo, or should it actively help reduce discrimination? If AI systems maintain the current gap between the autonomy of disabled and non-disabled social groups, calling them ‘enablers’ may be misleading. It might even be the opposite: AI might exclude particular user groups, including disabled individuals facing intersections of multiple forms of discrimination. Future research, accordingly, could move beyond the qualitative analysis of the enabling power of AI and investigate how much AI is improving the everyday lives of disabled individuals. To do this meaningfully, researchers and developers need to be clear about what exactly they are assessing: the potential of AI, its real-world impact, or its relative contribution compared to broader social progress. Making these distinctions is essential for building a deeper understanding of fairness and for designing AI systems that are genuinely inclusive.

Considering intersectional discrimination as one of the social forces that we studied, and given that we believe in more involvement of social forces in evaluating fair AI, this study can also provide a more practical contribution. We would, accordingly, coin the term *‘intersectional debiasing’* as an effort to include, if technically viable, not just persons with one underrepresented attribute but also persons who face intersections of discrimination in AI. While this concept still requires technical feasibility assessments, it provides a more socially inclusive mindset to shift algorithmic fairness to a new paradigm, one that integrates social justice principles and recognizes the multi-dimensional experiences of marginalized groups. Such a shift could meaningfully advance the discourse on inclusive and ethical AI.

References

Bondarouk, T., Ruël, H., 2004. Discourse analysis: making complex methodology simple, in: Leino, T., Saarinen, T., Klein, S. (Eds.), . Proceedings of the 12th European Conference on Information Systems (ECIS), Turku, Finland.

Braun, M., Wölfel, M., Renner, G., Menschik, C., 2020. Accessibility of Different Natural User Interfaces for People with Intellectual Disabilities, in: Proceedings - 2020 International Conference on Cyberworlds, CW 2020. Institute of Electrical and Electronics Engineers Inc., pp. 211–218. https://doi.org/10.1109/CW49994.2020.00041

Chiodo, S., 2022. Human autonomy, technological automation (and reverse). AI Soc 37, 39–48. https://doi.org/10.1007/s00146-021-01149-5

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madeling, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., Vayena, E., 2018. Ethical framework for a good AI society: opportunities, risks, principles, and recommendations. AI4People 28, 689–707. https://doi.org/10.1007/s11023-018-9482-5

Friedman, B., Hendry, D.G., 2019. Value Sensitive Design- Shaping Technology with Moral Imagination. The MIT Press, Cambridge, Massachusetts- London, England.

Haraway, D., 1991. A Cyborg Manifesto: Science, Technology, and Socialist-Feminism in the Late Twentieth Century, in: Simians, Cyborgs and Women: The Reinvention of Nature, Routledge. Routledge, New York, pp. 149–181.

Hertweck, C., Heitz, C., Loi, M., 2024. What's Distributive Justice Got to Do with It? Rethinking Algorithmic Fairness from the Perspective of Approximate Justice.

Hofmann, M., Kasnitz, D., Mankoff, J., Bennett, C.L., 2020. Living Disability Theory: Reflections on Access, Research, and Design, in: Proceedings of the 22nd International ACM SIGACCESS Conference on Computers and Accessibility. ACM, New York, NY, USA, pp. 1–13. https://doi.org/10.1145/3373625.3416996

Laitinen, A., Sahlgren, O., 2021. AI Systems and Respect for Human Autonomy. Front Artif Intell 4. https://doi.org/10.3389/frai.2021.705164

Latour, B., 2005. Reassembling the Social: An Introduction to Actor-Network-Theory. Oxford University Press, Oxford.

Lawson, A., Beckett, A.E., 2020. The social and human rights models of disability: towards a complementarity thesis. International Journal of Human Rights 1–32. https://doi.org/10.1080/13642987.2020.1783533

Lillywhite, A., Wolbring, G., 2023. Coverage of well-being within artificial intelligence, machine learning and robotics academic literature: the case of disabled people. AI Soc. https://doi.org/10.1007/s00146-023-01735-9

Lythreatis, S., Singh, S.K., El-Kassar, A.N., 2022. The digital divide: A review and future research agenda. Technol Forecast Soc Change 175. https://doi.org/10.1016/j.techfore.2021.121359

Mazera, M.S., Schneider, D.G., Padilha, M.I., Amadigi, F.R., Bruggmann, M.S., 2024. The perception of people with physical disabilities about exercising autonomy in a federal university. Texto e Contexto Enfermagem 33. https://doi.org/10.1590/1980-265X-TCE-2022-0194en

McLuhan, M., 1994. Understanding Media, The Extension of Man. MIT Press, London and New York.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A., 2019. A Survey on Bias and Fairness in Machine Learning.

Mengibar, A.C., 2015. Critical discourse analysis in the study of representation, identity politics and power relations: A multi-method approach. Communication and Society 28, 39–54. https://doi.org/10.15581/003.28.2.39-54

Mitra, S., 2006. The capability approach and disability. J Disabil Policy Stud. https://doi.org/10.1177/10442073060160040501

Moyà-Köhler, J., Domènech, M., 2022. Challenging 'The Hands of Technology': An Analysis of Independent Living for People with Intellectual Disabilities. Int J Environ Res Public Health 19. https://doi.org/10.3390/ijerph19031701

Nierling, L., Maia, M., 2020. Assistive technologies: Social barriers and socio-technical pathways. Societies 10. https://doi.org/10.3390/soc10020041

Noble, S.U., 2020. Algorithms of Oppression. New York University Press. https://doi.org/10.18574/nyu/9781479833641.001.0001

Pagano, T.P., Loureiro, R.B., Lisboa, F.V.N., Cruz, G.O.R., Peixoto, R.M., Guimarães, G.A. de S., Santos, L.L. dos, Araujo, M.M., Cruz, M., de Oliveira, E.L.S., Winkler, I., Nascimento, E.G.S., 2022. Bias and unfairness in machine learning models: a systematic literature review.

Park, S., Humphry, J., 2019. Exclusion by design: intersections of social, digital and data exclusion. Inf Commun Soc 22, 934–953. https://doi.org/10.1080/1369118X.2019.1606266

Prunkl, C., 2024. Human Autonomy at Risk? An Analysis of the Challenges from AI. Minds Mach (Dordr) 34. https://doi.org/10.1007/s11023-024-09665-1

Rawls, J., 1971. A Theory of Justice. Harvard University Press, Cambridge, MA.

Shams, R.A., Zowghi, D., Bano, M., 2023. AI and the quest for diversity and inclusion: a systematic literature review. AI and Ethics. https://doi.org/10.1007/s43681-023-00362-w

Shew, A., 2020. Ableism, Technoableism, and Future AI. IEEE Technology and Society Magazine 39, 40-50+85. https://doi.org/10.1109/MTS.2020.2967492

Timpe, K., 2019. Moral Ecology, Disabilities, and Human Agency špace 1pc 2018 Wade Memorial Lecture. Res Philosophica 96, 17–41. https://doi.org/10.11612/resphil.1741

Tsatsou, P., 2020. Digital inclusion of people with disabilities: a qualitative study of intra-disability diversity in the digital realm. Behaviour and Information Technology 39, 995–1010. https://doi.org/10.1080/0144929X.2019.1636136

Wall, J.D., Stahl, B.C., Salam, A.F., 2015. Critical discourse analysis as a review methodology: An empirical example. Communications of the Association for Information Systems 37, 257–285. https://doi.org/10.17705/1cais.03711

Whittaker, M., Alper, M., Bennett, C.L., Hendren, S., Kaziunas, L., Mills, M., Ringel Morris, M., Rankin, J., Rogers, E., Salas, M., Myers West, S., 2019. Disability, Bias, and AI (Workshop report by AI Now Institute). New York.

Wolbring, G., 2024. Workshop: how can research be more diverse and inclusive. Presentation at Institute of Technology Assessment and System Analysis (ITAS).

Wolbring, G., Nasir, L., 2024. Intersectionality of Disabled People through a Disability Studies, Ability-Based Studies, and Intersectional Pedagogy Lens: A Survey and a Scoping Review. Societies 14, 176. https://doi.org/10.3390/soc14090176

Annex 1. Title of the Entries as Empirical Data (Sorted by Organization Categories, Alphabetically)

Title		Author(S)	Organization	Entry Category	Publishing Date	Collecting Date	Link
1	Can ChatGPT Make the World More Accessible	Benjamin Roussey	Accessibility	Accessibility & Disability Advocacy	Apr 03, 2023	May 03, 2024	https://www.accessibility.com/blog/can-chatgpt-make-the-world-more-accessible?ref=disabilitydebrief.org
2	GPT-4 Image Recognition: An Absolute Game Changer in Accessibility	Aaron Preece	American Foundation of the Blind	Accessibility & Disability Advocacy	Feb 09, 2024	May 03, 2024	https://afb.org/blog/entry/gpt-4-image-recognition-accessibility?ref=disabilitydebrief.org
3	Microsoft Leverages Power of AI To Improve Accessibility for Disabled People	Sarah Sarsby	AT Today	Accessibility & Disability Advocacy	May 19, 2023	May 03, 2024	https://attoday.co.uk/microsoft-leverages-power-of-ai-to-improve-accessibility-for-disabled-people/?ref=disabilitydebrief.org
4	New GPT-4 Model Can Reportedly Describe Images Accurately	NA	boia (Bureau of Internet Accessibility)	Accessibility & Disability Advocacy	Apr 20, 2023	May 03, 2024	https://www.boia.org/blog/new-gpt-4-model-can-reportedly-describe-images-accurately?ref=disabilitydebrief.org
5	Artificial Intelligence Products for Disabled People	Krissie Barrick	Disability Charity: Scope UK	Accessibility & Disability Advocacy	NA	Sep 30, 2024	https://www.scope.org.uk/news-and-stories/artificial-intelligence-disabled-people
6	How Will AI Help Disabled People	Emma Purcell	Disability Horizons Shop	Accessibility & Disability Advocacy	NA	Sep 30, 2024	https://shop.disabilityhorizons.com/how-will-ai-help-disabled-people/
7	Using AI for Disability Inclusion	Kristina Treadwell	Disability:IN	Accessibility & Disability Advocacy	NA	Sep 30, 2024	https://disabilityin.org/business-case/using-ai-for-disability-inclusion/

Title		Author(S)	Organization	Entry Category	Publishing Date	Collecting Date	Link
8	Is AI A Risk or an Opportunity for Disability Rights	Shah Maitreya	European Network on Independent Living	Accessibility & Disability Advocacy	NA	Sep 30, 2024	https://enil.eu/is-ai-a-risk-or-an-opportunity-for-disability-rights/
9	AI for Accessibility: Opportunities and Challenges	Cindy Bennett, Shari Trewin	Equal Entry	Accessibility & Disability Advocacy	Mar 28, 2023	May 03, 2024	https://equalentry.com/ai-for-accessibility-opportunities-and-challenges/?ref=disabilitydebrief.org
10	Real AI Solutions for Accessibility Challenges	Kevin Berg	Equal Entry	Accessibility & Disability Advocacy	Sep 26, 2023	May 03, 2024	https://equalentry.com/real-ai-solutions-for-accessibility-challenges/?ref=disabilitydebrief.org
11	How AI Needs to be Redesigned for People with Disabilities	Sam Proulx	Fable	Accessibility & Disability Advocacy	NA	Sep 30, 2024	https://makeitfable.com/article/ai-and-analytics-people-with-disabilities/
12	AI for Disability Inclusion: Friend or Foe	NA	Get Skilled Access	Accessibility & Disability Advocacy	NA	Sep 30, 2024	https://getskilledaccess.com.au/blog/ai-for-disability-inclusion/
13	AI For All: Why Disability Inclusion Is Vital to the Future of Artificial Intelligence	NA	Scope	Accessibility & Disability Advocacy	May 16, 2024	Sep 30, 2024	https://www.linkedin.com/pulse/ai-all-why-disability-inclusion-vital-future-artificial-intelligence-z59ne/
14	Fairness of AI for People with Disabilities: Problem Analysis and Interdisciplinary Collaboration	Jason J.G. White	SIG Access	Accessibility & Disability Advocacy	Oct, 2019	Sep 30, 2024	https://www.sigaccess.org/newsletter/2019-10/white.html
15	Three Ways AI Supports People with Disabilities in the Workplace	NA	Verbit.ai via Accessibility	Accessibility & Disability Advocacy	Mar 9, 2023	Sep 30, 2024	https://www.accessibility.com/blog/three-ways-ai-supports-people-with-disabilities-in-the-workplace
16	No, 'AI' Will Not Fix Accessibility	Adrian Roselli	Adrian Roselli	Blog & NGO & Independent Authors	Sep 08, 2024	May 03, 2024	https://adrianroselli.com/2023/06/no-ai-will-not-fix-accessibility.html?ref=disabilitydebrief.org

Title		Author(S)	Organization	Entry Category	Publishing Date	Collecting Date	Link
17	Ableism and ChatGPT: Why People Fear It Versus Why They Should Fear It	Mich Ciurria	Blog of the APA	Blog & NGO & Independent Authors	Mar 30, 2023	May 03, 2024	https://blog.apaonline.org/2023/03/30/ableism-and-chatgpt-why-people-fear-it-versus-why-they-should-fear-it/?ref=disabilitydebrief.org
18	Nothing About Us, Without Us: Disability Justice and AI	Kenrya Rankin	Mozilla Foundation	Blog & NGO & Independent Authors	July 09, 2024	Sep 30, 2024	https://foundation.mozilla.org/en/blog/disability-justice-and-ai/
19	Adventures with BeMyAI	Léonie Watson	Tink	Blog & NGO & Independent Authors	Aug 17, 2023	May 03, 2024	https://tink.uk/adventures-with-bemyai/?ref=disabilitydebrief.org
20	Disability, Accessibility, and AI - Towards Data Science	Stephanie Kirmer	Towards Data Science	Blog & NGO & Independent Authors	Sep 16, 2024	Sep 30, 2024	https://towardsdatascience.com/disability-accessibility-and-ai-0d5ab06ec140
21	Report – To Reduce Disability Bias in Technology, Start with Disability Data	Ariana Aboulafia, Miranda Bogen	Center for Democracy and Technology (cdt)	Blog & NGO & Independent Authors	July 25, 2024	Oct 01, 2024	https://cdt.org/insights/report-to-reduce-disability-bias-in-technology-start-with-disability-data/
22	Equally AI Releases ChatGPT-Powered Report on Web Accessibility Websites in the US, Urges Business Leaders to Prioritize Inclusivity	Kathy Berardi	PRWeb	Business, Tech & Market Analysis	Mar 29, 2023	May 03, 2024	https://www.prweb.com/releases/equally-ai-releases-chatgpt-powered-report-on-web-accessibility-websites-in-the-us-urges-business-leaders-to-prioritize-inclusivity-809668774.html
23	Artificial Intelligence Is Dangerous for Disabled People at Work: 4 Takeaways for Developers and Buyers	Nancy Doyle	Forbes	Business, Tech & Market Analysis	Oct 11, 2022	May 03, 2024	https://www.forbes.com/sites/drnancydoyle/2022/10/11/artificial-intelligence-is-dangerous-for-disabled-people-at-work-4-takeaways-for-developers-and-buyers/?ref=disabilitydebrief.org
24	Disability Data Alarmingly Absent from AI Algorithmic Tools, Report Suggests	Gus Alexiou	Forbes	Business, Tech & Market Analysis	Aug 06, 2024	Sep 30, 2024	https://www.forbes.com/sites/gusalexiou/2024/08/06/disability-data-alarmingly-absent-from-ai-algorithmic-tools-report-suggests/

Title		Author(S)	Organization	Entry Category	Publishing Date	Collecting Date	Link
25	Empowering Individuals with Disabilities Through AI Technology	Tyler Weitzman	Forbes	Business, Tech & Market Analysis	Jun 16, 2023	Sep 30, 2024	https://www.forbes.com/councils/forbesbusinesscouncil/2023/06/16/empowering-individuals-with-disabilities-through-ai-technology/
26	Envision Adds ChatGPT AI Sight Assistance to Its Smart Glasses for the Blind	Gus Alexiou	Forbes	Business, Tech & Market Analysis	Apr 30, 2023	May 03, 2024	https://www.forbes.com/sites/gusalexiou/2023/04/30/envision-adds-chatgpt-ai-sight-assistance-to-its-smart-glasses-for-the-blind/?ref=disabilitydebrief.org
27	How AI Can Improve the Lives of People with Disabilities	NA	Smart Click	Business, Tech & Market Analysis	NA	Sep 30, 2024	https://smartclick.ai/articles/how-ai-can-improve-the-lives-of-people-with-disabilities/
28	How AI Is Advancing Assistive Technology	Mary K. Pratt	Tech Target	Business, Tech & Market Analysis	Jan 22,2024	May 03, 2024	https://www.techtarget.com/searchenterpriseai/tip/How-AI-is-advancing-assistive-technology?ref=disabilitydebrief.org
29	Be My Eyes Announces New Tool Powered by OpenAI's GPT-4 to Improve Accessibility for People Who are Blind or Have Low-Vision	NA	Business Wire	News & Media Outlet	Mar 14, 2023	Sep 30, 2024	https://www.businesswire.com/news/home/20230314005425/en/Be-My-Eyes-Announces-New-Tool-Powered-by-OpenAI0.000000E+002https://en.yna.co.kr/view/AEN20230317004500315?ref=disabilitydebrief.org
30	'We Don't Want To Leave People Behind': AI Is Helping Disabled People in Surprising New Ways	Clare Duffy	CNN Business	News & Media Outlet	July 08, 2024	Sep 30, 2024	https://edition.cnn.com/2024/07/08/tech/ai-assistive-technology-disabilities/index.html
31	Can AI Be Used to Help People with Disabilities? Experts Say Yes, With The 'Right Data Set'	Irelyne Lavery	Global News	News & Media Outlet	Jan 29, 2023	Sep 30, 2024	https://globalnews.ca/news/9440455/artificial-intelligence-disability/
32	How Ableist Algorithms Dominate Digital Spaces	John Loeppky	IT Pro	News & Media Outlet	Feb 20, 2023	May 03, 2024	https://www.itpro.com/technology/artificial-intelligence-ai/370064/how-ableist-algorithms-dominate-digital-spaces?ref=disabilitydebrief.org

Title		Author(S)	Organization	Entry Category	Publishing Date	Collecting Date	Link
33	AI Revolution: Paralyzed Woman 'Speaks' via Digital Avatar	Robin Marks	Neuroscience News	News & Media Outlet	Aug 23, 2023	May 03, 2024	https://neurosciencenews.com/ai-bci-voice-recreation-23810/?ref=disabilitydebrief.org
34	GPT-4's New Capabilities Power A 'Virtual Volunteer' For the Visually Impaired	Devin Coldewey	TechCrunch	News & Media Outlet	Mar 14, 2023	May 03, 2024	https://techcrunch.com/2023/03/14/gpt-4s-first-app-is-a-virtual-volunteer-for-the-visually-impaired/?ref=disabilitydebrief.org&guccounter=1&guce_referrer=aHR0cHM6Ly93d3cuZGlzYWJpbGl0eWRlYnJpZWYub3JnL2xpYnJhcnkvdG9waWMtZGlnaXRhbGFpLw&guce_referrer_sig=AQAAAD3WCZAJp_0mo-DGordWLn8SLwPdSOMU3_Hl8Xr0rPAbq8AbpYseabU6zPyuYix4kwE46w0kXbtX9wLW1l8ae15kXzGNsjxlQscUWHNQQAkMO4a1l-7pg7aPkqjFxeFnt1AtXJ5g3VT37ilrQBUdtM1Uk5xJaGdA8t95LRt_64CI
35	Why AI Fairness Conversations Must Include Disabled People	Eileen O'Grady	The Harvard Gazette	News & Media Outlet	Apr 03, 2024	Sep 30, 2024	https://news.harvard.edu/gazette/story/2024/04/why-ai-fairness-conversations-must-include-disabled-people/
36	Common AI Language Models Show Bias Against People with Disabilities: Study	Gianna Melillo	The Hill: Changing America	News & Media Outlet	Oct 14, 2022	May 03, 2024	https://thehill.com/changing-america/respect/diversity-inclusion/3688507-common-ai-language-models-show-bias-against-people-with-disabilities-study/?ref=disabilitydebrief.org
37	Book Review: 'Against Technoableism,' by Ashley Shew	Andrew Leland	The New York Times	News & Media Outlet	Sep 19, 2023	Sep 30, 2024	https://www.nytimes.com/2023/09/19/books/review/against-technoableism-ashley-shew.html?utm_source=cmpgn_news&utm_medium=email&utm_campaign=vtAdvUnirelClipReportsCMP_weeklysept212023

	Title	Author(S)	Organization	Entry Category	Publishing Date	Collecting Date	Link
38	Automating Ableism	S.E. Smith	The Verge	News & Media Outlet	Feb 14, 2024	Sep 30, 2024	https://www.theverge.com/24066641/disability-ableism-ai-census-qalys
39	More Equal Opportunities: How AI Fosters an Inclusive Working World	NA	acatech (National Academy for Science and Engineering)	Public Institution & International Organization	July 06, 2023	Sep 30, 2024	https://en.acatech.de/allgemein/how-ai-fosters-an-inclusive-working-world/
40	Artificial Intelligence and Its Impact on the Human Rights of Persons with Disabilities	Jerneja Turin,	European Network of National Human Rights Institutions	Public Institution & International Organization	Dec 03, 2023	Sep 30, 2024	https://ennhri.org/news-and-blog/artificial-intelligence-and-its-impact-on-the-human-rights-of-persons-with-disabilities/
41	Can AI Improve the Lives of Persons with Disabilities	Klaus Hoeckner	Futurium	Public Institution & International Organization	Feb 21, 2019	Sep 30, 2024	https://futurium.ec.europa.eu/en/european-ai-alliance/blog/can-ai-improve-lives-persons-disabilities
42	AI Act and Disability-Centred Policy: How Can We Stop Perpetuating Social Exclusion?	Yonah Welker	OECD	Public Institution & International Organization	May 17, 2023	Sep 30, 2024	https://oecd.ai/en/wonk/eu-ai-act-disabilities
43	Humanity Should Get the Best From AI, Not the Worst	NA	UN Human Rights	Public Institution & International Organization	May 09, 2022	May 03, 2024	https://www.ohchr.org/en/stories/2022/05/humanity-should-get-best-ai-not-worst-un-disability-rights-expert?ref=disabilitydebrief.org
44	The AI Revolution: Is it a Game Changer for Disability Inclusion?	Hudoykul Hafizov	UNDP Uzbekistan	Public Institution & International Organization	July 18, 2024	May 03, 2024	https://www.undp.org/uzbekistan/blog/ai-revolution-it-game-changer-disability-inclusion

Title		Author(S)	Organization	Entry Category	Publishing Date	Collecting Date	Link
45	Algorithms, Artificial Intelligence, and Disability Discrimination in Hiring	NA	US Department of Justice Civil Rights Divition	Public Institution & International Organization	May 12, 2022	Sep 30, 2024	https://www.ada.gov/resources/ai-guidance/
46	Generative AI Holds Great Potential for Those with Disabilities - But It Needs Policy to Shape It	Yonah Welker	World Economic Forum	Public Institution & International Organization	Nov 03, 2023	May 03, 2024	https://www.weforum.org/agenda/2023/11/generative-ai-holds-potential-disabilities/?ref=disabilitydebrief.org
47	How Cognitive Diversity in AI Can Help Close the Disability Inclusion Gap	Yonah Welker	World Economic Forum	Public Institution & International Organization	Apr 17, 2023	Oct 01, 2024	https://www.weforum.org/agenda/2023/04/how-cognitive-diversity-and-disability-centred-ai-can-improve-social-inclusion/?ref=disabilitydebrief.org
48	How Sovereign Funds Could Empower the Future of Assistive Technology and Disability AI	Yonah Welker	World Economic Forum	Public Institution & International Organization	Aug 15, 2023	Oct 01, 2024	https://www.weforum.org/agenda/2023/08/sovereign-funds-future-assistive-technology-disability-ai/?ref=disabilitydebrief.org

Annex 2. The Codes and Subcodes of the Qualitative Data Analysis