Towards a method for exploring meaningful explanations of algorithmic processes

Igor ter Halle[1], Pascal de Vries[1]

[1]Research group Digital Business & Society, Windesheim University of Applied Sciences, The Netherlands

DOI 10.3217/978-3-99161-033-5-006, CC BY 4.0

Abstract. In an era where algorithmic processes increasingly influence our daily lives, the need for comprehensible explanations of these processes is growing. This paper introduces the development of a methodological approach to explore the possibilities for meaningful explanations of algorithmic processes. Utilizing both theoretical frameworks and empirical case studies, this study aims to bridge the gap between algorithmic complexity and user understanding. The proposed method emphasizes transparency and accessibility, supporting policymakers and technology designers in creating more insightful and accountable technological applications. Key findings from this research highlight the critical role of interdisciplinary approaches in shaping effective explanation mechanisms, which are essential for fostering an ethically responsible integration of technology into society.

Introduction

This paper presents the first findings of a project. In this project, a methodology is developed to help governments to make algorithm use more explainable. In this paper, we first describe the Dutch context, then provide a brief overview of the existing literature and then describe the method used to arrive at a methodology that can be used to discuss explainability in concrete use cases around automated decision making.

1. Context

Dutch citizens generally exhibit moderate trust in their government. The Dutch are conservative in their confidence in political institutions, assigning on average a grade of six out of ten (Grimmelikhuijsen 2018). The average political trust in the Netherlands fluctuates over time, yet there is no discernible long-term downward or upward trend.

Political trust tends to correlate with major national developments such as economic crises or viral outbreaks, subsequently returning to its original level.

It is noteworthy that the variance in political trust has structurally increased over the past fifteen years, indicating a greater divergence in trust levels among the Dutch populace. Furthermore, disparities exist between different levels of political institutions; local political bodies tend to inspire greater trust than national (and international) counterparts. Recent investigations by the Dutch newspaper Trouw have also revealed a high level of trust in the police and judicial system within the Netherlands. However, when it comes to specific aspects of governance, such as the government's fulfillment of promises, transparency in policy formation, and equitable treatment of all citizens, the Dutch are significantly more critical and pessimistic. Process satisfaction emerges as a vital contributor to political trust.

Grimmelikhuijsen (2018) posits that many assume transparency to be beneficial for governmental trust. Nevertheless, his research suggests that while transparency serves multiple purposes well, it does not inherently enhance trust, particularly concerning political decision-making. The less politically oriented an organization is, the more trust it engenders through transparent operations. This is evidenced by studies on the judiciary and regulatory bodies, where transparency has been shown to positively influence trust (Grimmelikhuijsen, 2018).

The underlying premise of a transparent government is the notion that if governmental organizations demonstrate to citizens (and other stakeholders) the decision-making processes, including how decisions are made and their outcomes, trust in the government will naturally increase (Grimmelikhuijsen, 2012). Thus, transparency is employed in practice as a standard tool to elevate trust (Grimmelikhuijsen, 2013).

2. The introduction of an algorithm register

To enhance transparency concerning algorithms used by the government, an online Algorithm Register has been available since December 2022. Government agencies publish information about the algorithms they employ within this register. An Algorithm Register is a public database that provides detailed information about the algorithms utilized by an organization. The content of the register may vary, but it typically includes the objectives of the algorithm, the data it processes, its operational mechanisms, and its impact on decision-making processes. The register offers information on the purpose and impact of the algorithm, any conducted Data Protection Impact Assessment (DPIA), or an Impact Assessment on Human Rights and Algorithms (IAMA), along with the data sources used.

Currently, the filling of the Algorithm Register is voluntary, hence it has not yet gained significant traction among the intended audience. However, according to the Action Agenda for Value-Driven Digitalization, by 2025 all algorithms relevant to citizens are required to be included in the Algorithm Register, deviating only when explicitly permitted by law or justified considerations.

In essence, government organizations will be responsible for providing insight into algorithms, thereby establishing and managing an Algorithm Register. The Data Protection Authority (DPA), as the algorithmic regulator, generally oversees the use of algorithms. This oversight applies not only to government bodies but naturally extends to businesses and other organizations as well.

A notable aspect of this initiative, wherein algorithms are published in a register, is the expectation that online publication of written documentation serves as an appropriate form of transparency. In the Netherlands, this is achieved by publishing a written description of the algorithm in the register, which may be available as a downloadable document. In other cases, such as in France, proactive efforts have been made to reach less language-proficient individuals through videos and audio presentations (Lovelace Institute, 2021).

3. Technical transparency versus explainability

Within the realm of algorithmic transparency, the Ministry of Justice and Security distinguishes between 'technical transparency' and 'explainability'.

Technical transparency pertains to disclosing all technical details of an algorithm, including the source code. Explainability refers to elucidating the operation of the algorithm to the concerned citizen.

3.1. Technical Transparency

Technical transparency primarily aims to facilitate the auditing of algorithms (Court of Audit, 2021). Algorithmic auditing can take various forms (gov.uk, 2022), such as verifying documentation, testing algorithmic outcomes, or examining internal operations. Audits may be conducted by external entities, regulatory bodies, researchers, or other parties initiating an audit independently. Auditing serves to ensure internal assurance or verify compliance with legal standards. The scope and depth of audits will vary based on the risks, the context of algorithm use, and existing legal requirements.

3.2. Explainability

A common critique of decision-making algorithms is their resemblance to inscrutable black boxes (Selbst, 2018). Users, citizens, and even designers often do not comprehend how algorithms make decisions, making it impossible to trace their decision-making processes (Lima et al., 2022). The widespread deployment of influential decision-making algorithms has necessitated an understanding of their operation. Explainable artificial intelligence (XAI) is an academic field dedicated to enhancing people's understanding of decision-making algorithms, emerging from this necessity. Although XAI seems concerned with artificial intelligence (AI), much of its literature is also applicable to less complex algorithms, such as those currently published in the Algorithm Register.

As defined by Arrieta et al. (2020), an explainable system is "one that produces details or reasons to make its functioning clear or easy to understand" for a specific audience, whether they be users, designers, patients, citizens, or policymakers.

Many XAI papers view explainability primarily in terms of clarifying the (technical) system to make it less opaque. However, following de Bruijn et al. (2020), we consider explainability more as a socio-technical challenge that addresses both technology and social aspects together. The focus should be on the impact and building trust, not solely on overcoming opacity. Here, it is helpful to consider explainability in terms of mutual intelligibility, a concept from linguistics (Bloomfield, 1926). Languages are mutually intelligible if speakers of one language can understand speakers of another without significant difficulty or study. Mutual intelligibility is a continuum; there are degrees of intelligibility, not a stark division between intelligible and unintelligible.

4. Meaningful explanation

At the Dutch Ministry of the Interior, a standard has been established for the publication of algorithms in the algorithm register, differentiating between experts and citizens. However, the exact identity of these experts and the precise information they require is not entirely clear. Are they policy staff, developers, or others? Given the difficulty in defining this target audience, there is a risk that the register could become so comprehensive as to be less accessible to the citizens for whom it was initially intended, especially as citizens themselves are not a exactly a homogeneous group.

Therefore, with such a variety of target groups that the register could serve, different modes of explanation may be necessary. The nature of the explanation might depend on the complexity of the context in which (complex) algorithms will be used, the type of data involved, the intention and purpose of its use, and, consequently, to whom it should be explained.

In addition to the complexity of target groups, according to de Bruijn et al. (2020), the context in which the algorithm is explained should also be considered. De Bruijn et al. distinguish two axes (figure 1): the degree of politicization and the impact of the algorithm on the life of the citizen. If the algorithm to be explained relates to a politically sensitive topic, then trust in the explanation of the algorithm will likely be low. If the impact of an algorithmic decision on citizens is significant, it may lead to the politicization of the decision and challenges to the explanation. Thus, explaining algorithms in complex situations will not always enhance trust.

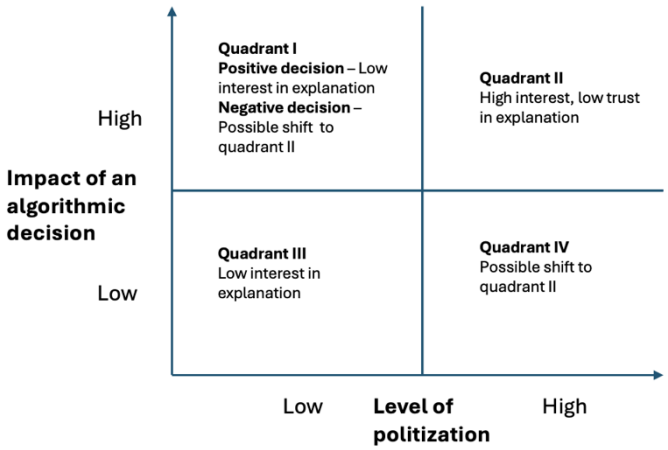

Figure 5. - Adapted from de Bruyn et al (2020)

Consequently, explaining the operation of an algorithm does not invariably lead to increased trust by citizens in the government or the deciding authority. It is therefore valuable to explore how to explain algorithms that have a high impact on the lives of citizens and a low to moderate degree of politicization to them.

A starting point for exploring potential strategies for explaining algorithms is the list of explanation strategies outlined by de Bruijn et al. (2020), based on the quadrant. These strategies include shifts from 1) explaining algorithms to explaining decisions, 2) from designing algorithms to co-creating/negotiating algorithms, 3) from explainable algorithms to explainable processes, 4) from an instrumental to an institutional approach, 5) from monopolistic algorithms and datasets to competing algorithms and datasets, 6) explaining the sensitivity to values of algorithms and how they have been addressed, particularly regarding gender, ethnicity, age, etc., and 7) from algorithms that replace professional decision-making to professionals who challenge algorithmic decision-making. As the challenges mentioned above are interconnected, a combination of strategies will typically be necessary.

5. Towards an approach for exploring meaningful explanations

To gain insight into how to meaningfully explain algorithms in practice, an approach is being developed to co-create actionable alternatives with professionals (to whom it may concerns). This approach looks for handles that professionals can use to provide contextual explanations about the algorithm and the context in which it has been used.

In the development of this approach, inspiration was sought from two methods that appear to be potentially suitable:

- Human-Centered Design (HCD)

- Guidance Ethics Approach (GEA)

HCD is frequently mentioned in the literature surrounding explainable artificial intelligence and is part of the approach (see, for example, Schoonderwoerd et al., 2021, Hall et al., 2019). The Guidance Ethics Approach is used within our Digital Business & Society research group to find concrete handles to apply technology in an ethically responsible manner (see https://ecp.nl/publicatie/guidance-ethics-approach for a full description of the approach).

5.1. Human centered design

Figure 2 presents a process flow diagram for a human-centered explanation design. In Human-Centered Design (HCD), three components are distinguished as crucial within the design process (see Schoonderwoerd et al., 2021, among others): domain analysis, requirements elicitation, and interaction design. Each component produces outcomes that serve as input for the next part.

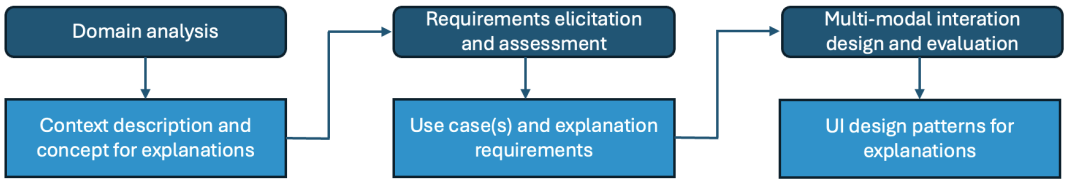

Figure 6. Flowchart explainable AI (Adapted from Schoonderwoerd et al., 2021)

5.1.1. Domain Analysis

A fundamental premise of all human-centered design approaches is to gain an understanding of the context of use (see, for example, Hall et al., 2019). The purpose of this insight is to determine whether and why explanations are needed and which information could be considered relevant in the context under examination.

The outcome of domain analysis is a description of the context in which a user seeks an explanation and an initial concept for the explanations based on the information that is relevant to end-users.

5.1.2. Requirements Elicitation

The objective here is to ascertain what kinds of explanations the system should be capable of providing. This process aims to outline a rich context (i.e., a use case or scenario) from which the target group's requirements can be identified (Maguire and Bevan, 2002). Wolf (2019) targets the development of usage scenarios where explanations are likely to be relevant (i.e., explanation scenarios).

5.1.3. Interaction Design

This phase's goal is to discover how the developed explanations can be effectively communicated. This includes selecting suitable modalities for presenting the information, typically involving a multimodal combination of visual and textual content (Holzinger et al., 2021).

6. Guidance ethics approach

The Guidance Ethics Approach (GEA) develops concrete action options for handling technology through structured dialogue with stakeholders within a sector or organization. From various user perspectives, the technological innovation under analysis is explored. The approach is employed to develop alternatives for AI applications.

It is also utilized in the development and usage of digital healthcare solutions. The method was developed in collaboration with Professor Peter-Paul Verbeek by the Platform for the Information Society and is described on the website begeleidingsethiek.nl (see https://ecp.nl/wp-content/uploads/2020/11/Guidance-ethics-approach.pdf for a summary in English).

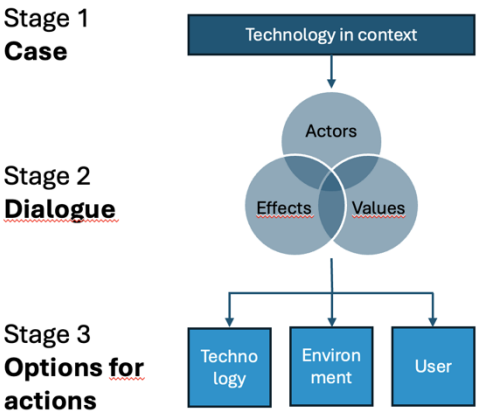

Figure 7. Guidance ethics approach (adapted from begeleidingsethiek.nl)

The approach involves a workshop where different stakeholders within an organization or sector engage in dialogue about the application of a specific technology within a specific context. The goal of the workshop is to jointly arrive at concrete action alternatives, ensuring that the technology discussed is embedded in the day-to-day operations within the organization or sector.

The workshop comprises three phases (see Figure 3). It begins with describing the technology and the concrete context in which it operates, focusing on a clear and comprehensible description without excessive jargon or technical details, making it understandable for an interested outsider.

In the second phase, the potential effects of deploying a technology in that context are explored. We seek to understand who is involved with the technology and which values are pertinent in daily practice. Ideally, the actual stakeholders would contribute to the dialogue. If not, all stakeholders can participate, representatives may think from their perspective. Distinguishing various effects can aid in acquiring a rich and realistic view of technology use. There are always multiple values associated with technology; in most cases, various values are significant. Like the effects, the process begins with an open inventory, followed by determining which values are deemed most relevant.

In the final phase, action options are formulated. Three types of action options are distinguished: from the technology ('ethics by design'), from the context ('ethics in context'), and from the user ('ethics in use').

7. Towards the meaningful explanation approach

After exploring both approaches through brainstorming, it has been decided to divide the workshop into two phases. The first phase explores the case from the participants' perspectives. In this exploration, the participants examine the algorithm to be explained from the perspectives of the actors involved with the algorithm. All these actors have expectations, and the outcome of the algorithm (such as whether or not a housing allowance is granted) has implications for the involved actor. Thus, the first phase explores the case and the various perspectives of the actors. Questions to be addressed in the first phase include:

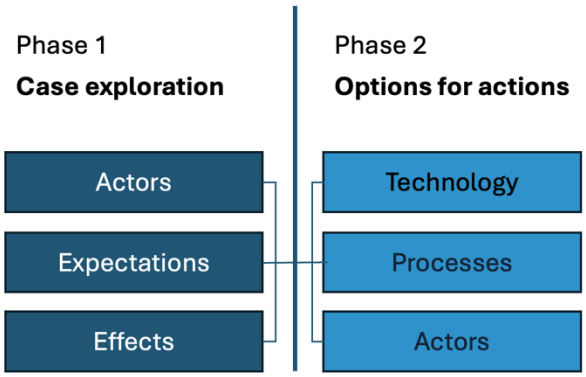

Figure 8. Meaningful explanation approach (1st version)

- Which actors play a role in explaining algorithms and procedures?
- What role do these actors play? What are their expectations?
- To whom should what be explained?
- What are the consequences of (in)comprehensible explanations for these actors?
- When is an actor satisfied?

Building on this exploration, the second phase investigates action alternatives from the perspective of technology ('transparency by design'). This includes personalized explanations, the use of multimedia, etc. Possibilities are also explored from the perspective of the process in which the algorithm is implemented ('transparency in process'). Consider, for example, the role a helpdesk might play or relevant parties in the process (such as housing associations, advocacy groups, etc.). Finally, the workshop explores ways to improve explanations from the various actors' viewpoints ('transparency in use').

Working with the meaningful explanation approach clarifies where meaning resides in the process of explanation. Following this, alternatives for action can be explored to achieve that meaning.

7.1. Testing the method

The method was first tested in a round table dialogue with participants from a department of the Dutch tax authorities. Six participants who are involved in an algorithm on the allocation of a rent allowance. Each participant had a different roles in the discussed algorithm (developers, policy advisors, helpdesk etc.). In the first fase of the dialogue the participants started from the perspective of different stakeholders such as users,

developers, policy makers and decision makers. After the dialogue, the participants generated options for action.

In order to prevent the dialogue from becoming mainly a theoretical exploration, we used one concrete algorithm to discuss the practical use of an algorithm: what is it? what does it do? who has anything to do with it? and what are the experiences with regard to explainability and transparency. A lot was discussed during the dialogue session. Entering into the dialogue in itself is already an exercise in transparency and explainability among colleagues.

Conducting a dialogue is not that easy. The participants were all personally invited by a colleague based on their specific areas of contact with the subject. Care was taken to ensure a mixed group of participants (in terms of expertise, gender, age, etc.)

A conclusion emerged from that process; participants indicated that it is good to talk to each other (from different practices) about a concrete algortihm. The structure of the dialogue session was used very loosely. The dialogue touched on all subjects and therefore it became a very natural dialogue. This does cause some difficulties for the reporting because the content can no longer be placed so well within the original framework.

After first test of the method, it is evident that the exchange of values around meaning has not yet been given a place, although there is a need for it. In the Guidance Ethics Approach (GEA), values are explored and identified in the dialogue between different perspectives of actors as represented by the participants. In the meaningful explanation approach, values are named from the perspective of the explainer. This clarifies what is alive among the participants and which values are recognized. The approach, therefore, works with different perspectives, but these are introduced by the explainer. Whether this aligns with the perspective of the explainees has not been definitively established.

Participation in an GEA-dialogue shows that introducing different perspectives of ownership provides the opportunity to discuss the process of the case study. We have not actively questioned the process of the algorithm, and it did not emerge organically, even though the selection of speakers took into account the constructivist nature of an algorithm (Seaver, 2018). Explaining should, after all, be part of the algorithmic process. This deserves attention next time to become part of the approach.

In the future, we aim to further employ this instrument in discussions about meaningful explainability within government organizations and refine it so that it becomes a tool to discern which alternatives for action contribute to a meaningful explanation of algorithmic processes.

References

Alexander, C. (1977). A pattern language: towns, buildings, construction. Oxford University Press.

Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, pp.82-115.

Bloomfield, L. (1926). A set of postulates for the science of language. Language, 2(3), pp.153-164.

De Bruijn, H., Warnier, M., & Janssen, M. (2022). The perils and pitfalls of explainable AI: Strategies for explaining algorithmic decision-making. Government Information Quarterly, 39(2), 101666.

Grimmelikhuijsen, S. (2012). Linking transparency, knowledge and citizen trust in government: An experiment. International Review of Administrative Sciences, 78(1), pp.50-73.

Grimmelikhuijsen, S. (2013). Meer openbaarheid, meer vertrouwen?. B en M: Tijdschrift voor Beleid, Politiek en Maatschappij, 40(4), pp.451-455.

Grimmelikhuijsen, S. (2018). Van gegeven naar verdiend gezag: Hoe kan transparantere rechtspraak (blijvend) bijdragen aan legitimiteit?. Rechtstreeks, 15(2), pp.13-35.

Hall, M., Harborne, D., Tomsett, R., Galetic, V., Quintana-Amate, S., Nottle, A., & Preece, A. (2019). A systematic method to understand requirements for explainable AI (XAI) systems. In Proceedings of the IJCAI Workshop on eXplainable Artificial Intelligence (XAI 2019), Macau, China, Vol. 11.

Holzinger, A., Malle, B., Saranti, A., & Pfeifer, B. (2021). Towards multi-modal causability with graph neural networks enabling information fusion for explainable AI. Information Fusion, 71, pp.28-37.

Lima, G., Grgić-Hlača, N., Jeong, J. K., & Cha, M. (2022). The conflict between explainable and accountable decision-making algorithms. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, pp.2103-2113.

Maguire, M., & Bevan, N. (2002). User requirements analysis: a review of supporting methods. In IFIP World Computer Congress, TC 13, Boston, MA: Springer US, pp.133-148.

Schoonderwoerd, T.A., Jorritsma, W., Neerincx, M.A., & Van Den Bosch, K. (2021). Human-centered XAI: Developing design patterns for explanations of clinical decision support systems. International Journal of Human-Computer Studies, 154, 102684.

Selbst, A.D., & Barocas, S. (2018). The intuitive appeal of explainable machines. Fordham Law Review, 87, 1085.

Seaver, N. (2018). What should an anthropology of algorithms do?. Cultural Anthropology, 33(3), pp.375-385.

Verbeek, P.P., & Tijink, D. (2020). Guidance Ethics Approach: an ethical dialogue about technology with perspective on actions. ECP | Platform voor de Informatie Samenleving.