The perils and pitfalls of explainable AI

Strategies for explaining algorithmic decision-making

de Bruijn, Hans; Warnier, Martijn; Janssen, Marijn

DOI
10.1016/j.giq.2021.101666

Publication date
2022

Document Version
Final published version

Published in
Government Information Quarterly

Government Information Quarterly 39 (2022) 101666

Contents lists available at ScienceDirect

Government Information Quarterly

journal homepage: www.elsevier.com/locate/govinf

The perils and pitfalls of explainable AI: Strategies for explaining algorithmic decision-making

Hans de Bruijn, Martijn Warnier, Marijn Janssen *

Delft University of Technology, Delft, the Netherlands

ARTICLE INFO

Keywords:
Artificial intelligence
XAI
Algorithms
Computational intelligence
Data-driven decision
Socio-tech
Transparency
Accountability
Trust
E-government

ABSTRACT

Governments look at explainable artificial intelligence's (XAI) potential to tackle the criticisms of the opaqueness of algorithmic decision-making with AI. Although XAI is appealing as a solution for automated decisions, the wicked nature of the challenges governments face complicates the use of XAI. Wickedness means that the facts that define a problem are ambiguous and that there is no consensus on the normative criteria for solving this problem. In such a situation, the use of algorithms can result in distrust. Whereas there is much research advancing XAI technology, the focus of this paper is on strategies for explainability. Three illustrative cases are used to show that explainable, data-driven decisions are often not perceived as objective by the public. The context might raise strong incentives to contest and distrust the explanation of AI, and as a consequence, fierce resistance from society is encountered. To overcome the inherent problems of XAI, decisions-specific strategies are proposed to lead to societal acceptance of AI-based decisions. We suggest strategies to embrace explainable decisions and processes, co-create decisions with societal actors, move away from an instrumental to an institutional approach, use competing and value-sensitive algorithms, and mobilize the tacit knowledge of professionals

1. Introduction

Public organizations increasingly use artificial intelligence (AI) for automating and supporting their decision-making, and there has been a steady increase in publications on this topic (Sousa, Melo, Bermejo, Farias, & Gomes, 2019. AI can enable new services for citizens, businesses, and public agencies (Kankanhalli, Charalabidis, & Mellouli, 2019) or automate existing ones. The deployment of new electronic services would likely increase government effectiveness and efficiency (Bertot, Estevez, & Janowski, 2016). But also other public values like accountability, transparency, equality, privacy and security, sustainability, and interoperability should be given attention when designing AI for public use (Kankanhalli et al., 2019).

The use of AI encounters many challenges (Sun & Medaglia, 2019). The decisions made by autonomous computational algorithms can severely impact both individuals and organizations (Brauneis & Goodman, 2018) and influence the power balance between governments, businesses, and citizens. As algorithms become increasingly autonomous and invisible, it is becoming harder to see and explain them (Janssen & Kuk, 2016). An algorithmic society might be too opaque to be accountable for its behavior (Brauneis, Goodman, & Tech., 2018). What decisions computational algorithms make, based on what information and how they make these decisions, should be explained to the public.

XAI has the potential to explain the working of AI to the general public. Although there has been much research on AI in the public sector context, *explainable artificial intelligence* (XAI) has been given less attention. XAI is a field based on the idea that advice given by expert systems would be more acceptable to humans if the advice could be explained to them (Swartout & Moore, 1993; Swartout, Paris, & Moore, 1991). XAI contrasts with opaque, black-box approaches that often cannot explain where a decision comes from or how it is justified. The more complex AI models are built, the more accurate these are, but the explainability of their working might be lost (Xu et al., 2019). In this paper, XAI is defined as *the extent to which AI outcomes are insightful for the general public*.

Explainability is an intuitively appealing concept but is hard to realize. Belle and Papantonis (2021) provide four suggestions for creating explainability, including explanation by simplification, describing the contribution of each feature to the decisions, explaining

* Corresponding author.
E-mail address: m.f.w.h.a.janssen@tudelft.nl (M. Janssen).

https://doi.org/10.1016/j.giq.2021.101666
Received 7 March 2021; Received in revised form 5 December 2021; Accepted 19 December 2021
Available online 30 December 2021

an instance instead of in general, and using graphical visualization methods for explanations. At the same time, they also discuss the complexity of realizing such suggestions. Simplifications might not be correct, features can be interrelated, local explanations can fail to provide the complete picture, and graphical visualization requires assumptions about data that might not necessarily be true.

Explainability is assumed to create transparency and trust in AI. Although trust might be affected in different ways than expected, situational factors also affect trust (Bannister & Connolly, 2011b). Transparency can both increase or decrease trust (Bannister & Connolly, 2011a). In a similar vein, XAI might either increase or decrease trust. Hence the nature of explainability should be understood better, and strategies are needed for creating trust in XAI.

In this paper, we derive the challenges of XAI and develop strategies for overcoming these challenges. This paper is structured as follows. In Section 2, we provide an overview of the concept of XAI followed by three illustrative case studies that demonstrate the challenges of XAI in Section 3. Section 4 explains why XAI is so challenging to realize and use in government. In Section 5, the relationships between XAI, transparency and trust are presented, followed by a proposal for strategies on how to deal with explainability in algorithmic decision-making. Finally, in Section 6, conclusions are drawn.

2. Explainable AI literature

The XAI research field has its origin in the early 90s in the field of expert systems (Xu et al., 2019). Pioneers such as Swartout and Moore reasoned that advice-giving by expert systems would be more acceptable to humans if the expert system could explain why it gave a particular advice (Swartout et al., 1991; Swartout & Moore, 1993). Expert systems are based on a large collection of rules that try to capture the knowledge of an expert. Hence the name 'expert system' is used. Rules are typically described in the form of implications, from which new conclusions can be derived if specific premises hold. An explanation consists of a trace of the application of rules with conforming conclusions and premises. An explanation looks like "the system came to this diagnosis because it applied these rules in this order to these initial symptoms, thereby concluding that the patient has this sickness". Such explanations, named trace explanations, were the first type of explanations.

At a later stage, more explanation types emerged, including justification, strategy, and terminological explanations (Gregor & Benbasat, 1999). In essence, these techniques are trace explanations enriched with more domain knowledge. In this way, explanations become easier to interpret. For example, by explaining domain-specific terminology or adding justifications to the rules used by the expert system. Different explanation types are typically combined to explain the outcome of expert systems. Similar techniques are used for explaining the behavior of other systems, including training systems (Harbers, 2011), medical support systems (Holzinger, Biemann, Pattichis, & Kell, 2017), legal support systems (Doshi-Velez et al., 2017), and educational systems (Conati, Porayska-Pomsta, & Mavrikis, 2018).

All the above systems have in common is that they are based on an underlying *symbolic representation.* Although intended for processing by machines, symbolic systems use languages (symbols), which are understandable by humans and that humans can use to verify the reasoning. For decision-making, the logic to arrive at a decision is simulating human reasoning, for instance, through the rules "if .. then .. ". Such rules are often in the form of computer code built and understood by people. Furthermore, while an expert may understand symbolic systems' logic, the logic might not be easy to understand for non-experts (Preece, Harborne, Braines, Tomsett, & Chakraborty, 2018). Therefore, the focus of XAI is often on *interpreting* whether the results are correct. Hence, instead of XAI, sometimes the term *interpretable machine learning* is used to explain and present model behavior in understandable terms to humans (Du, Liu, & Hu, 2019).

In contrast to symbolic systems, *non-symbolic* systems, including popular machine learning models such as deep learning, arrive at decisions by connecting the inputs with the outputs (LeCun, Bengio, & Hinton, 2015). Such systems produce the correct answer without executing the logic to arrive at this answer. The logic cannot be grasped directly by humans. These systems are not directly programmed and constructed by humans, but rather are sophisticated statistical models that 'learn' by being trained on large amounts of data using machine learning techniques, such as neural networks or deep learning (Jaeger, 2016). The internal representation of these non-symbolic AI systems does not contain a collection of human-readable rules but instead a collection of non-linear correlations. Whereas the logic in symbolic systems is verifiable by experts, non-symbolic systems are more cumbersome, as translation steps are required to make them understandable to humans. Therefore, only post-hoc analysis can be conducted to verify the results. Many of these algorithms are continuously trained with new data and learn from their own decisions (Jordan & Mitchell, 2015). The continuous training or self-learning of algorithms will also influence the explanations. Hence, explanations also need to be continuously updated.

XAI envisions the construction of a symbolic, human-understandable model automatically from the non-symbolic, statistical machine-learned model. This would make the model interpretable, and hence, it would be relatively straightforward to explain the outcomes of the system. There is a large body of academic literature on the interpretation and explanation of machine-learned models (Biran & Cotton, 2017; Du et al., 2019; Montavon, Samek, & Müller, 2018; Samek, Wiegand, & Müller, 2017). Techniques used for explaining machine-learned models range from relatively straightforward sensitivity analysis (Zhang & Wallace, 2015) to highly complex machine learning techniques, such as Taylor decomposition (Montavon, Lapuschkin, Binder, Samek, & Müller, 2017) and layer-wise relevance propagation (Bach et al., 2015). Advanced models built to explain machine learning models' outcomes deploy machine learning techniques themselves. The explanations generated by such models are typically not easy to understand and require interpretation by experts (Du et al., 2019). As a consequence, the explanations, at least to some extent, are open to interpretation. The more complex the situation, the more challenging it is to explain the results.

Some researchers have distinguished between model-centric explanations (an explanation of the AI model itself, as focused on above) for general information-sharing and broader accountability purposes; and more subject-centric explanations (explanations of how a particular decision has impacted a particular individual or group). An example of the latter are so-called recourse algorithms that look at the possible harm caused by decisions and reverse such decisions. If harmful outcomes are generated in a range of counterfactual scenarios, then these need to be reversed. Such an algorithm generates candidate changes to the variables that would reverse an algorithm's decision(Venkatasubramanian & Alfano, 2020). In this way, the harmful outcomes of AI-based decisions are removed. Note that, while a promising research direction, such techniques still need to be developed further in order to be useful in practice and deployed on a large scale.

Based on the state of the art literature, we can conclude that XAI is a promising research field that is quickly developing, but that there are not yet widely available techniques that can be easily deployed to provide unambiguous explanations of the underlying AI models. This is especially true for the non-symbolic machine learning-based models, that remain hard to explain and interpret, especially to non-experts.

Due to the outlined challenges to understanding an explanation, the term "meaningful" is sometimes included in practice (Pedreschi et al., 2019). *Meaningful* refers to systems providing explanations that are understandable to individual users. As there are many types of users, different explanations might be required. What meaningful explainability looks like will likely depend on the complexity of the context in which AI will be used, the type of data used, the intent and purpose for its use, and whom it should be explained to. To understand this prospect better, three illustrative case studies are presented in the next section.

3. The complexity of XAI in practice

The three cases were selected because they are illustrative of different failures to explain AI-based decisions. In this paper, we take three prominent cases situated in three different countries and use them to analyze and illustrate the challenges of XAI in the next section. The three cases are selected because they are illustrative of the failure to explain AI-based decisions.

3.1. Risks analysis of fraud in the social domain in the Netherlands

System Risk Identification (SyRI) is a fraud detection system based on the integration of personal data from several databases controlled and maintained by public agencies in the Netherlands (Public Interest Litigation Project, 2020). This project dismantles the traditional silos in which multiple agencies store data. Almost any kind of data was allowed to be shared for the broad goal of detecting fraud using AI. By connecting all governmental data about citizens, potentially anybody became the subject of this type of analysis. Due to SyRI, entire neighborhoods were identified as potentially fraudulent. Black-boxed AI models were used to determine who does or does not come into view of the enforcement and investigation services, e.g., neighborhoods with expensive cars and low incomes. Although the motive of fraud detection is not contested by society, the use of AI to carry out such detection can be problematic. Even though explaining how data and algorithms are used to select certain groups is possible, society contested the approach. Anybody can be of interest to the investigators, which violates the principle that people are considered innocent until proven guilty.

3.2. Immigration in the UK

In the UK, AI is used to make decisions on whether immigrants are allowed to enter the country (McDonald, 2019). Although the final decision remains in the hand of the civil servants, AI is used to determine which cases require more scrutiny. Questions started to be asked about how AI affects the immigration policy and the rights of immigrants. One of the reasons for such questions is the fear that people belonging to different AI-created groups will be treated in different ways. This could result in “fast lanes” that would lead to “speedy boarding for white people” (McDonald, 2019). Although this might not be viewed as a racial bias by some, the idea of having persons added to a group in which their chance of getting a visa differs from that of other groups is against the expectation of non-discrimination and equal treatment. This challenges the public values of equal opportunity and non-discrimination. The counter-argument used is that AI does not make decisions, and the final decision is in the hands of the human caseworkers. Although such caseworkers should consider the outcomes of AI as a suggestion, their own critical thinking might be affected or reduced. They might not even have the abilities or be given the freedom to contest the outcomes of the AI-generated decision.

3.3. Re-offending of criminals

In the USA, thousands of court cases are fed into algorithms to predict whether a defendant will commit a new crime or fail to return to court. As with the previous cases, the projects were initiated by the government, however in this case the tool was developed by a commercial company with the aim to give judges the most objective information available to make decisions about prisoners' risk of re-offending (Lynn, 2018). Defendants are given a risk score that is presented to the judge. Judges use these risk scores to make pre-trial decisions on defendants. Some persons have praised the system for ensuring that dangerous persons are kept in jail, whereas harmless persons go free. Besides, the system does not use race, gender, employment, or living place to avoid decision bias based on race, gender, or appearance and to arrive at more objective outcomes. Nevertheless, a study showed that the system predicted that black defendants pose a higher risk of recidivism (Larson, Mattu, Kirchner, & Angwin, 2016). Although race is not included in the data fed to the AI system, racial bias is introduced by utilizing other data sources. Similar to the previous case, the system was also criticized for substituting and removing the judges' critical thinking and nudging them towards biased recommendations. The judges can ignore the AI-generated recommendations if they consider them wrong, but they should justify their decision to deviate instead of that the AI-generated decision is motivated. In addition, the ability of the AI system to learn from its own decisions has both pros and cons. By learning from their own mistakes, these mistakes can be prevented in the future. At the same time, this raises the question if the AI system is sufficiently mature for taking into production and use. Finally, the algorithm used in this case is proprietary and not open for scrutiny by the public.

All three cases aim to address societal problems, reduce civil servant bias, and make more objective decisions by using data and algorithms. Despite that the working of the algorithms could be explained, up to a point, in the first two cases, AI may result in outcomes that do not meet the commonly accepted public values of non-discrimination, equal treatment and, judges' independence. Accurate predicting is never fully possible using these types of algorithms. Even if the results can be explained, the outcomes can be contested and mistrusted. Although the decisions might be perceived as objective, they are objected by society.

4. Challenges of XAI

XAI is an intuitively appealing concept, as explanations are something that is desired. However, society's norms and values can be translated differently and what is acceptable differs between societies. Even within a single society, norms might be different. That is why deliberation is deemed to be necessary for policy-making (Gerston, 2004). Policy-making is a process that focuses on reaching a consensus about the norms. Although the reasons why explainable AI does not fulfill its intended purpose are intertwined, we describe these reasons using the following seven main challenges of XAI summarized in Table 1. The challenges are multi-faceted and intertwined.

The *first* objection against XAI is obvious. XAI often focuses on ensuring that solutions are understood by the public (Swartout et al., 1991; Swartout & Moore, 1993). XAI assumes a certain level of expertise of the public, however, many persons will simply lack that expertise and cannot assess whether an AI-based decision is fair or just (Du et al., 2019). Tools and instruments can be used to abstract and explain decisions, but this results in a further deviation of how the actual AI-based decision is taken, and the simplification might be contestable (Belle &

Table 1
Summary of the challenges of XAI.

Challenge	Explanation
1. Lack of expertise	Most persons will lack the expertise to understand the explanation and assess the fairness of the decision.
2. Contested explanations	Experts explaining algorithms also make biased and inherently disputable choices.
3. Dynamics of data and decisions	Data and decisions change over time, and therefore explanations change.
4. Interference of algorithms	Often there is a whole chain of activities to collect and process data from various types of sources, and many, often different kinds of algorithms are used.
5. Context-dependency	Algorithms cannot be explained at a general level, as outcomes might be different per individual.
6. Wicked nature of the problems addressed	Wicked problems are ill-structured, are ambiguous by nature and can be solved in different ways. Algorithms provide one answer that is contestable and changes over time.
7. Causality is not used for making decisions	If the causality is explained between inputs and outputs, this does not mean that the algorithm uses that causality to arrive at a decision. Furthermore, the explanation of causality might change over time.

Papantonis, 2021). Besides, XAI is not about making a single decision but about a large number of decisions. The emerging 'fast lane' in the immigration case study shows that although a single decision might be accepted, but the overall outcome is contested. Whereas an individual case can be explained, the overall effect cannot.

Second, explaining AI-based decisions is not a neutral process. For example, in the re-offending case, data about use race, gender, employment, and living place are left out to avoid bias, racial profiling, and reinforcing historical discrimination. Nevertheless, this resulted in false objectivity as these characteristics can be reflected by other variables. Decisions are based on a complex interaction between data and possibly multiple algorithms, and in the process of deriving the conceptual 'translation' that a broader audience can understand, the translator also makes inherently disputable (biased) choices. For instance, political preferences might be reflected by the explanation. In turn, others might arrive at alternative explanations, and the explanation given might be contested. Furthermore, no single explanation would do the work for all decisions made by an algorithm.

The *third* reason is that the algorithm is all too often not static. Algorithms learn from their own decisions and by incorporating new data (Jordan & Mitchell, 2015). Hence algorithms are dynamic and change over time. The learning part can be contested. In the re-offending case, the argument was that the AI system would learn and improve from its mistakes. At the same time, this argument suggests that AI does not work properly and makes mistakes. This raises the question of whether it is fair to use the AI system at all. The more dynamic an algorithm's context is, the more challenging XAI will be. These dynamics result in today's explanation becoming obsolete tomorrow. In addition, it might even be unclear for experts what exactly changed. You cannot explain what you do not know.

The *fourth* reason is that algorithms interfere with each other. A decision made by an algorithm is influenced by the, inherently biased, data collected from different sources (Janssen, van der Voort, & Wahyudi, 2017), by the other (learning) algorithms (Bozdag, 2013; Janssen & Kuk, 2016), and by the context where the algorithm is deployed (Janowski, 2015). This is shown, for example, in the case of risk analysis in the social domain. Adding more data and using a variety of algorithms hinders explainability and might introduce bias or racial profiling, as happened in the re-offending case. Furthermore, the social domain typically depends on the tacit knowledge of the situation at hand. Tacit knowledge is not codified and therefore cannot be taken into account by an algorithm. The variety of algorithms hampers explainability – for the same reason as under the third objection: the inference is neither knowable nor explainable.

Fifth, algorithmic decisions might have different consequences for different individuals. In some cases, XAI will only be meaningful if the AI-based decision can be explained at an individual level. Put differently, the explanation of AI is very much context-dependent (Selbst et al., 2019). This makes it hard to explain the working of the algorithms in general and to explain their different outcomes. Transparency might be meaningless if algorithms make decisions that influence an individual's life (Bozdag, 2013). Furthermore, equal treatment is an important public value, but individuals are different and are treated differently. Individualized XAI means that the four previous objections manifest in almost infinite variations. Take the re-offending case. If a defendant wants to know how the algorithm makes a decision, it is hard to compare that decision with that of others , as their situation differs. The logic to compare with other cases is unclear. Furthermore, segmentation between criminals and non-criminals is fair for society, but a segmentation between low and high incomes is not considered fair. In the risk analysis in the social domain, the United Nations (UN) commented on the use of AI for the segmentation stating that treating poor people differently from the rich is not acceptable (Lynn, 2018). Yet, at the heart of many AI applications is some kind of segmentation algorithm.

The *sixth* reason is that algorithms are often used to address so-called wicked problems in which traditional rule-based systems are not appropriate or easy to use, and algorithms search for the underlying, non-visible patterns (Jordan & Mitchell, 2015). Problems are wicked because the relevant facts are ambiguous, and the actors involved disagree on the question of how to value a problem morally. Wickedness implies that there is no single right problem definition and no single right solution. Explaining using the traditional cause-effect relationship is not applicable to wicked problems (Rittel & Webber, 1973) and, as such, cannot be used for explaining them. Furthermore, there might be many solutions to deal with these types of problems, and the algorithms might only provide one possible solution. Again, this results in contesting the outcomes. Furthermore, these 'solutions' typically change over time for wicked problems (Rittel & Webber, 1973). Per definition, it is impossible to explain something ambiguous.

Finally, The ability to explain the causality behind an AI-based decision does not mean that the AI system actually uses this causality and that the actual relationship between inputs and outputs might be different. In all three cases, the AI systems were black-boxed systems whose inner working was opaque. Even if we could 'explain' an algorithm – how do we know (can we verify) that this is the same one that is actually used? This can only be verified in practice by having regular audits and inspecting how the algorithm works. Different explanations might be needed for the same algorithms over time. For the public, changing explanations over time is neither understandable nor acceptable.

5. Trust in explainable AI

XAI should be conducive to the trust and societal acceptance of AI-based decisions. XAI should result in legitimate (unbiased) decisions and trust in the effectiveness and fairness of such decisions. XAI is aimed at creating transparent and clear decision-making processes and decisions, which in turn should result in trust in decisions. However, the relationship between transparency and trust is often more complicated than assumed (Bannister & Connolly, 2011a). Transparency might influence trust in such a way that consumers do not purchase, donors do not give, shareholders do not invest, or governments are not trusted (Auger, 2014, p. 327). As such, even the best XAI might not result in the desired trust. People who do not know how AI works will not have sufficient confidence in it and thus will not use or approve its outcomes.

In addition to these limitations of XAI, there is the societal context in which algorithms are used. This context can also have a major impact on the explainability of AI. Finally, the possible societal impact influences the explainability. The content and impact can be used for classifying the situations for deploying XAI, as shown in Fig. 1.

- *The context of using algorithms might be less or more politicized*. Politicization refers to the level of possible conflict and the chance of having contested outcomes. Contested outcomes refer to issues in which the public has opposing opinions, like abortion and immigration, often dependent on their political preference. Therefore, the decisions made by AI are also politicized and show the preference of those in power.
- *The impact of an AI-based decision might also be high or low*. For example, AI-based decisions on law enforcement priorities are potentially high-impact decisions, determining whether citizens will be faced with the police. Other decisions, like a fine for speeding, might have a low impact.

When having high levels of politicization and high impact (Quadrant II), the public will be very critical of endeavors to explain AI – the AI-based decision has a high impact and is operating in a highly politicized context. That means that XAI will likely be conflict-generating, and parties will have a different view on what constitutes the right decision and if a decision is perceived as fair. The latter might depend on their political preference. The public will be interested in the AI, underlying

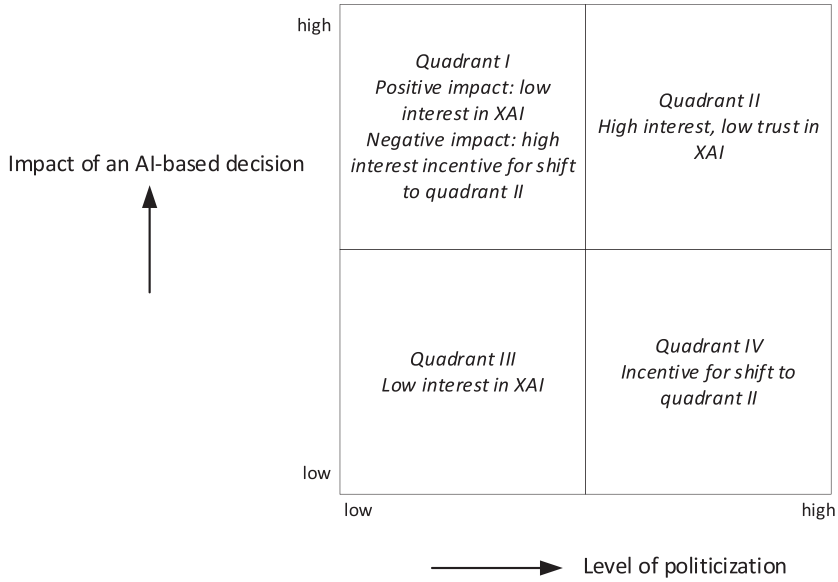

Fig. 1. Contextualizing XAI.

the decision that affects them, and will likely distrust the explanation. They will always find an argument for their distrust, given the limitations of XAI listed in the preceding section and their political preferences.

The opposite situation can be found when there is a low level of politicization and low impact (Quadrant III). Likely, this means that people will not be interested in an explanation of the AI-based decision. Decisions simply do not matter and have limited consequences. These types of decisions might be relatively simple, and black-boxed AI solutions might not be needed. Rule-based systems that are easy to explain are sufficient, and XAI might not be needed at all.

When there is a highly politicized environment, even an AI-based decision with a low impact can be perceived as problematic (Quadrant IV). The public might be suspicious of every decision and question the outcomes. Due to the contestable nature of the decisions, the decisions might be perceived as having a high impact, which results in a shift towards high interest and low trust in AI. The more politicization, the higher the chance that a decision will be perceived as having a high impact. Hence, this type of decision might shift to the second quadrant.

Finally, in a low politicized context having a high impact (Quadrant I), AI-based decision outcomes will probably be accepted more easily than in a highly politicized context. However, a difference between decisions having a positive impact (e.g., a firm receiving a subsidy, an immigrant-receiving a residence permit) and a negative impact (e.g., the subsidy or the permit are denied) need to be made. If a high-impact decision is negative, the person affected negatively might be inclined to file a complaint and even politicize the decision by stating that there is something wrong with the subsidy policy or that the system is broken. In summary, when the impact of a decision is high, negative decisions carry incentives for politicization.

In summary, XAI is always applied in a specific context – and this context might have strong incentives to contest and distrust the explanation of AI, irrespective of XAI's quality. If the level of politicization is high, the trust in XAI will probably be low, even for low-impact decisions. If the impact of a decision is high, there might be incentives to politicize the issue and to challenge XAI. All three cases presented in this paper can be positioned in the high politicized and high impact quadrant (II), although their initial position might be in other quadrants. Due to the outcomes' possible negative or high impact, their position was shifted to the high politicized and impact quadrant. Strategies for trusted XAI need to take into account the high politicized and high impact nature.

6. Strategies for trusted XAI

XAI has fundamental limitations for use in complex situations and will not always provide an acceptable explanation or improve trust. Hence we derive a series of strategies for policy-makers that might contribute to more legitimacy and trust (Bruijn, Janssen, & Warnier, 2020). The strategies are presented in a 'from, to' frame to show the shift in emphasis needed. The strategies can be both an alternative, or an addition, to XAI. The strategies suggest a shift away from an instrumentation approach focused on the AI algorithm towards an approach focused on creating legitimacy and trust in decisions. There is no one-to-one connection between the challenges and the strategies. The strategies do not solve the challenges necessarily, as much is dependent on how these are realized. Furthermore, the strategies should be used in concert.

6.1. Strategy 1: From explaining AI to explaining decisions produced using AI

We could shift our attention from explaining AI to explaining the decision supported by AI. A decision might be fully or partially based on AI, but in any case, decision-makers should be able to explain why a decision has been made. When decision-makers have this burden of proof, there might be an incentive to scrutinize the algorithms used or deviate from AI-based decision-making critically. It might make them decide not to rely on AI or on a particular type of AI exclusively.

Take the example of re-offending criminals. A judge must always argue which considerations have led to the verdict - in a way that is transparent to the litigating parties and to society. This obligation to explain the final decision can be an incentive to reflect on the role of the algorithm critically. After all, the mere fact that an algorithm has a specific outcome is not sufficient to argue a decision - either what the algorithm does is properly explained, or the outcome of the algorithm is wholly or partially ignored.

6.2. Strategy 2: From designing algorithms to negotiated algorithms

In some cases, algorithms can be more authoritative if they are not designed by experts only. Instead, an approach of co-creation with the public and interested parties can be taken. The main choices algorithms are based upon can be discussed and published. The parties involved can try to find consensus about, for example, the variables that are taken into account by an algorithm or the scope of an algorithm – what decisions an

algorithm should make and should not make and how humans should remain in control. A process like this results in *negotiated algorithms* in which every stakeholder has its say, and a consensus needs to be reached.

6.3. Strategy 3: From explainable algorithms to explainable processes

Closely related to the idea that algorithms can be explained is that the design process also can be explained. Transparency then refers to questions like who will be involved, who will have what role, what are the main issues that will be debated, how will parties deal with dissensus and uncertainties, how will they make their decisions. Not all discussions need to be documented in detail, but only the relevant processes that lead to decisions and the argumentation why decisions were taken. The attention shifts from making algorithms explainable to making the process of creating algorithms explainable.

The case study on re-offending criminals can serve as an example. A great deal of scientific expertise is available on crime, recidivism, conditions under which recidivism occurs, bias in detection and sentencing, etc. When various experts are involved in the development of an algorithm, this can contribute to a "negotiated" algorithm (strategy 2) and more awareness of the limitations and risks of an algorithm. A metaphor for the latter is the instruction leaflet for a medicine. Experts can be involved in developing algorithms and drawing up a leaflet for an algorithm: an overview of the risks and instructions on how the algorithm should or should not be used. If an algorithm is negotiated, the next step might be to design the process of negotiations (strategy 3). The transparency of this process can contribute to the acceptance of the negotiated algorithm.

6.4. Strategy 4: From an instrumental to an institutional approach

The value of XAI is often translated into tools or instruments that are conducive to more transparency. However, transparency also requires an institutional approach – the development of 'rules of the game' for dealing with AI. Institution-building can comprise setting up organizational structures that facilitate the development of these rules of the game. Examples are the establishment of regulators with authority to scrutinize and audit algorithms and to develop regulation or, within organizations, review committees that are positioned as countervailing powers of developers and users of algorithms.

Consider the case of risk analysis of fraud with social benefits. The world of social security has a well-developed institutional structure. There are, for example, professional benefit agencies, client interest groups, scientists and experts, and there is social advocacy. The higher the degree of institutionalization, the easier it is to design an institutional structure that can be used as a countervailing power in the design and use of algorithms. A continuous critical look from this countervailing power can be conducive to the right use of algorithms.

6.5. Strategy 5: From monopolistic algorithms and datasets to competing algorithms and datasets

Using a metaphor from the world of economics, organizations often employ monopolistic algorithms and monopolistic datasets to a lesser extent. They develop one algorithm or one family of algorithms and use these algorithms to base their decisions upon. The transparency of AI-based decision-making can be enhanced by deliberately using competing algorithms and datasets. Only if competing algorithms that are trained on independently collected datasets result in more or less the same decision, it might be reasonable to assume that this is a correct decision. If competing algorithms provide different decisions, a human decision-maker should take over. Although simultaneous failure of independently-built and -operating algorithms is less likely, a false feeling of trust might be created as multiple algorithms and data might all be wrong. Suppose courts use competing algorithms and these different algorithms result in different outcomes. This will probably be conducive to a critical reflection on the algorithms.

6.6. Strategy 6: From algorithms to value-sensitive algorithms

AI-based decision-making can reinforce deeply rooted biases, and therefore result in morally wrong decisions. When designing algorithms, the parties involved can take certain key values into account. One should aim to design the algorithm in such a way that data that might result in biases or discrimination (e.g., age, gender, race) is ignored and verify whether these undesired variables have an impact on the proposed decisions of the algorithm (Du et al., 2019). Furthermore, tests can be conducted if humans are treated in the same manner. This 'value sensitive' design (Friedman, 1996) of algorithms incentivizes the parties involved to be transparent about what values they want to safeguard and how these values are guaranteed.

6.7. Strategy 7: From algorithms replacing professional decision-making to professionals challenging algorithmic decision-making

There is a classic tension between analytical decision-making based on facts and figures, and intuitive decision-making of professionals based on their tacit knowledge. Both types of decision-making have their strengths and weaknesses. There is the risk that with the emergence of AI, intuitive decision-making will be replaced by predominantly analytical decision-making. Also, if professionals are replaced, then their tacit knowledge will be lost. They often have deep insight into the nature of societal problems and what should be taken into account. It might be a strategy to make AI-based decisions *and* ask professionals to make decisions based on their tacit knowledge. Decisions made by AI and humans can be compared for reliability and accuracy, and facilitate mutual learning.

In the migration example, the important question is which cases require more scrutiny than others. For the migration case, this strategy would mean that the algorithm and professionals will assess a number of cases. The question is then whether the judgments diverge - where they do, further analysis may reveal why this is the case, and lessons may be drawn for the design and use of algorithms.

There is little guidance below on which strategies should be put together and which shouldn't. The essence of the challenges presented in Table 1 is that the complexity and ambiguity of algorithms make it impossible to explain them in a simple and unambiguous manner. There is no single best strategy for dealing with the challenges, as such a portfolio of strategies should be employed. If, for example, there is a negotiated algorithm (strategy 2), there are institutional checks and balances (strategy 4), and there are regular opportunities for professionals to challenge algorithmic decision-making (strategy 7), then the combination of these strategies can contribute to greater trust in algorithms. Some of the strategies might complement each other, like strategy 2 and 6, in which citizen representatives are used to sensitize algorithm design to fundamental values. Strategy 1 and 3 look at different aspects, with the former focusing on decisions but not design, and the latter focusing on the design process. Each of the strategies has its own shortcomings and might not solve the challenge of XAI. The strategies might be time-consuming and resource-intensive. Also, the strategies might not result in the desired outcomes. Nevertheless, the portfolio of strategies can contribute to the legitimacy of the use of algorithms. Obviously, the XAI strategies may not be effective for the problems they were not designed to tackle. The strategies might not be able to tackle the problem in cases where XAI might not be a good solution for automating decision-making. Diagnosing the situation at hand should always be the first step before jumping into solutions.

7. Discussion and conclusions

XAI faces many challenges when used for consequential decision-

making by governments. Often XAI is approached as a technical problem in which better algorithms can facilitate explanation. This view neglects the wicked nature of the problems for which XAI is used. What might initially look like a simple problem for a public body to explain the working of an algorithm to the public is often far more complicated. The public lacks the expertise, explanations might be contested, explanations might change over time or differ per case, various data sources and algorithms are used, the working of the algorithms does not reflect the explanation, and the problems are ambiguous and can be tackled in different ways. Apart from these challenges, the socio-political context in which XAI is used might create strong incentives to distrust the explanations of AI, irrespective of the quality of the explanation. In particular, if there are high levels of politicization and the decisions' impact is high. Surprisingly, data-driven AI is often employed for automating highly politicized situations in which the decisions have a high impact, as these might be hard to automate in another way.

The challenges show the need to broaden the view on XAI beyond merely taking an instrumental view. XAI should be approached as a socio-technical challenge in which both technology and social aspects are addressed in concert. The focus should be on the impact and on creating trust and not only on overcoming the opaqueness. The strategies include shifts from 1) explaining AI to explaining decisions 2) from designing algorithms to 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) from algorithms to value-sensitive algorithms and 7) from algorithms replacing professional decision-making to professionals challenging algorithmic decision-making. As the challenges are multi-faceted and interrelated, a combination of strategies will be typically needed.

In further research, each strategy can be expanded and tested in practice. Strategies for creating legitimacy and trust in XAI are derived for use by policy-makers, but they can be used as a future research agenda on XAI. This research is explorative in nature and consequently has several limitations. The work is based on three illustrative case studies. In other case studies, new challenges might appear. The strategies might not be exhaustive and can be extended in further research. Also, the strategies are not tested in practice and we did not investigate how the strategies can be employed to realize successful outcomes. The XAI research agenda should be broadened and become more than opening an algorithmic black box. The actual use of XAI should be analyzed, its positive and negative impact evaluated, and socio-technical aspects captured.

In this paper, the focus was primarily on the algorithms. Outside the scope of the paper is the malfunctioning of the software infrastructure, human mistakes, the use of (non-)trusted technology and open-source software. These elements can influence the XAI outcomes and should be tackled as well. Finally, we recommend investigating the relationship between XAI, openness, transparency, accountability and explainability. These concepts are interrelated and can be unraveled in further research.

Marijn Janssen is a Full Professor in ICT & Governance in the Technology, Policy and Management Faculty of Delft University of Technology. His research is focused on ICT-architecting and design science in situations in which multiple public and private organizations need to collaborate, in which ICT plays an enabling role, there are various ways to proceed, and socio-technical solutions are constrained by organizational realities and political wishes. Marijn was nominated in 2018 and 2019 by Apolitical as one of the 100 most influential people in the Digital Government worldwide https://apolitical.co/lists/digital-government-world100. He has published over 600 refereed publications, his google h-score is 72, having over 20 K citations. More information: www.tbm.tudelft.nl/marijnj.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was partially supported by TAILOR, a project funded by EU Horizon 2020 research and innovation programme under GA No 952215. A preliminary version of this paper was published in the Dutch language.

References

Auger, G. A. (2014). Trust me, trust me not: An experimental analysis of the effect of transparency on organizations. *Journal of Public Relations Research, 26*(4), 325–343.
Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K. R., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS One, 10*(7), Article e0130140.
Bannister, F., & Connolly, R. (2011a). The trouble with transparency: A critical review of openness in e-government. *Policy & Internet, 3*(1), Article 8.
Bannister, F., & Connolly, R. (2011b). Trust and transformational government: A proposed framework for research. *Government Information Quarterly, 28*(2), 137–147.
Belle, V., & Papantonis, I. (2021). Principles and practice of explainable machine learning. *Frontiers in big Data, 4*, 688969. https://doi.org/10.3389/fdata.2021.688969
Bertot, J., Estevez, E., & Janowski, T. (2016). Universal and contextualized public services: Digital public service innovation framework. *Government Information Quarterly, 33*(2), 211–222.
Biran, O., & Cotton, C. (2017). Explanation and justification in machine learning: A survey. In *Paper presented at the In IJCAI-17 workshop on explainable AI (XAI)*.
Bozdag, E. (2013). Bias in algorithmic filtering and personalization. *Ethics and Information Technology, 15*(3), 209–227.
Brauneis, R., & Goodman, E. P. (2018). Algorithmic transparency for the smart city. *Yale JL & Tech., 20*, 103.
Bruijn, H.d., Janssen, M., & Warnier, M. (2020). Transparantie en Explainable Artificial Intelligence: beperkingen en strategieën. *Bestuurskunde, 29*(4), 21–29.
Conati, C., Porayska-Pomsta, K., & Mavrikis, M. (2018). AI in education needs interpretable machine learning: Lessons from open learner modelling. *arXiv preprint*. arXiv:1807.00154.
Doshi-Velez, F., Kortz, M., Budish, R., Bavitz, C., Gershman, S., O'Brien, D., & Wood, A. (2017). Accountability of AI under the law: The role of explanation. *arXiv preprint*. arXiv:1711.01134.
Du, M., Liu, N., & Hu, X. (2019). Techniques for interpretable machine learning. *Communications of the ACM, 63*(1), 68–77.
Friedman, B. (1996). Value-sensitive design. *Interactions, 3*(6), 16–23.
Gerston, L. N. (2004). *Public policy making: Process and principle*. M.E. Sharpe.
Gregor, S., & Benbasat, I. (1999). Explanation from intelligent systems: Theoretical foundations and implications for practice. *MIS Quarterly, 23*(4), 497–530.
Harbers, M. (2011). Explaining agent behavior in virtual training. In *Doctoral Disseration Utrecht University*.
Holzinger, A., Biemann, C., Pattichis, C., & Kell, D. B. (2017). What do we need to build explainable AI systems for the medical domain? *arXiv preprint*. arXiv:1712.09923.
Jaeger, H. (2016). Artificial intelligence: Deep neural reasoning. *Nature, 538*(7626), 467.
Janowski, T. (2015). Digital government evolution: From transformation to contextualization. *Government Information Quarterly, 32*(3), 221–236. https://doi.org/10.1016/j.giq.2015.07.001
Janssen, M., & Kuk, G. (2016). The challenges and limits of big data algorithms in technocratic governance. *Government Information Quarterly, 33*(3), 371–377.
Janssen, M., Van der Voort, H., & Wahyudi, A. (2017). Factors influencing big data decision-making quality. *Journal of business research, 70*(1), 338–345.
Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science, 349*(6245), 255–260.
Kankanhalli, A., Charalabidis, Y., & Mellouli, S. (2019). IoT and AI for smart government: A research agenda. *Government Information Quarterly, 36*(2), 304–309. https://doi.org/10.1016/j.giq.2019.02.003
Larson, J., Mattu, S., Kirchner, L., & Angwin, J. (2016). *How We Analyzed the COMPAS Recidivism Algorithm*. Propubica, 23 May 2016. Retrieved from https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.
LeCun, Y., Bengio, Y., & Hinton, G. J.n. (2015). *Deep learning., 521*(7553), 436–444.
Lynn, B. (2018). *Judges Now Using Artificial Intelligence to Rule on Prisoners*. Science & Technology, 7 February 2019. Retrieved from https://learningenglish.voanews.com/a/ai-used-by-judges-to-rule-on-prisoners/4236134.html.
McDonald, H. (2019). AI system for granting UK visas is biased, rights groups claim. *The Gaurdian, 29*(10), 29 Oct 2019. Retrieved from https://amp.theguardian.com/uk-news/2019/oct/29/ai-system-for-granting-uk-visas-is-biased-rights-groups-claim.
Montavon, G., Lapuschkin, S., Binder, A., Samek, W., & Müller, K. R. (2017). Explaining nonlinear classification decisions with deep Taylor decomposition. *Pattern Recognition, 65*, 211–222.
Montavon, G., Samek, W., & Müller, K. R. (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing, 73*, 1–15.

Pedreschi, D., Giannotti, F., Guidotti, R., Monreale, A., Ruggieri, S., & Turini, F. (2019). Meaningful explanations of black box AI decision systems. In *Paper presented at the Proceedings of the AAAI conference on artificial intelligence*.

Preece, A., Harborne, D., Braines, D., Tomsett, R., & Chakraborty, S. (2018). Stakeholders in explainable AI. *arXiv preprint*. arXiv:1810.00184.

Public Interest Litigation Project. (2020). System Risk Indication (SyRI). Retrieved from https://pilpnjcm.nl/en/dossiers/profiling-and-syri/.

Rittel, H. W. J., & Webber, M. M. (1973). Dilemmas in a general theory of planning. *Policy Sciences, 4*(2), 155–169.

Samek, W., Wiegand, T., & Müller, K. R. (2017). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint*. arXiv:1708.08296.

Selbst, A. D., Boyd, D., Friedler, S. A., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems [Paper presentation]. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM. https://doi.org/10.1145/3287560.3287598.

Sousa, W. G.d., Melo, E. R. P.d., Bermejo, P. H. D. S., Farias, R. A. S., & Gomes, A. O. (2019). How and where is artificial intelligence in the public sector going? A literature review and research agenda. *Government Information Quarterly, 36*(4), Article 101392. https://doi.org/10.1016/j.giq.2019.07.004

Sun, T. Q., & Medaglia, R. (2019). Mapping the challenges of artificial intelligence in the public sector: Evidence from public healthcare. *Government Information Quarterly, 36*(2), 368–383. https://doi.org/10.1016/j.giq.2018.09.008

Swartout, W. R., & Moore, J. D. (1993). Explanation in second generation expert systems. In *Second generation expert systems* (pp. 543–585). Springer.

Swartout, W. R., Paris, C., & Moore, J. (1991). Explanations in knowledge systems: Design for explainable expert systems. *IEEE Intelligent Systems, 6*(3), 58–64.

Venkatasubramanian, S., & Alfano, M. (2020). The philosophical basis of algorithmic recourse. In *Paper presented at the Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*.

Xu, F., Uszkoreit, H., Du, Y., Fan, W., Zhao, D., & Zhu, J. (2019). Explainable AI: A brief survey on history, research areas, approaches and challenges. In *Paper presented at the CCF international conference on natural language processing and Chinese computing*.

Zhang, Y., & Wallace, B. (2015). A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. *arXiv preprint*. arXiv:1510.03820.

Hans de Bruijn is professor of governance at the Faculty of Technology, Policy and Management Delft University of Technology in The Netherlands and visiting professor at the Department of Management, Economics and Industrial Engineering at Politecnico di Milano, Italy. He studied law and political science and obtained his doctorate in public administration. He conducts research into governance issues at the intersection of technology and society. This often focuses on the "wicked" nature of many technological developments, which are also embedded in networks of interdependencies between the main political and social actors involved. Application areas include energy, transport, IT and water. Recent books are "Management in Networks" (Routledge: London 2018, together with Ernst ten Heuvelhof) and The Art of Political Framing (Amsterdam: Amsterdam University Press 2019).

Martijn Warnier is full Professor of Complex System Design and chair of the Multi-Actor Department of Delft University of Technology, the Netherlands. He holds a PhD degree in Computer Science from the Radboud University Nijmegen (2006). In his research Martijn studies complex emergent behavior in and of socio-technical systems, specifically of power and ICT infrastructure and the combination thereof. He employs multi-paradigm simulation modeling to study system aspects, such as robustness, resilience, efficiency and reliability and designs adaptive interventions that use self-organization techniques to improve the performance of socio-technical systems on these and other aspects such as empowerment, security and privacy of end-users.

Marijn Janssen is a Full Professor in ICT & Governance in the Technology, Policy and Management Faculty of Delft University of Technology. His research is focused on ICT-architecting and design science in situations in which multiple public and private organizations need to collaborate, in which ICT plays an enabling role, there are various ways to proceed, and socio-technical solutions are constrained by organizational realities and political wishes. Marijn was nominated in 2018 and 2019 by Apolitical as one of the 100 most influential people in the Digital Government worldwide https://apolitical.co/lists/digital-government-world100. He has published over 600 refereed publications, his google h-score is 72, having over 21K citations. More information: https://www.tbm.tudelft.nl/marijnj.