Original Manuscript

Social Science Computer Review
1-16
© The Author(s) 2020
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0894439320980118
journals.sagepub.com/home/ssc
SAGE

Will Algorithms Blind People? The Effect of Explainable AI and Decision-Makers' Experience on AI-supported Decision-Making in Government

Marijn Janssen[1], Martijn Hartog[1], Ricardo Matheus[1], Aaron Yi Ding[1], and George Kuk[2]

Abstract

Computational artificial intelligence (AI) algorithms are increasingly used to support decision making by governments. Yet algorithms often remain opaque to the decision makers and devoid of clear explanations for the decisions made. In this study, we used an experimental approach to compare decision making in three situations: humans making decisions (1) without any support of algorithms, (2) supported by business rules (BR), and (3) supported by machine learning (ML). Participants were asked to make the correct decisions given various scenarios, while BR and ML algorithms could provide correct or incorrect suggestions to the decision maker. This enabled us to evaluate whether the participants were able to understand the limitations of BR and ML. The experiment shows that algorithms help decision makers to make more correct decisions. The findings suggest that explainable AI combined with experience helps them detect incorrect suggestions made by algorithms. However, even experienced persons were not able to identify all mistakes. Ensuring the ability to understand and traceback decisions are not sufficient for avoiding making incorrect decisions. The findings imply that algorithms should be adopted with care and that selecting the appropriate algorithms for supporting decisions and training of decision makers are key factors in increasing accountability and transparency.

Keywords

AI, artificial intelligence, decision making, e-government, algorithmic governance, transparency, accountability, XAI, experiment, data-driven government

[1] Delft University of Technology, the Netherlands
[2] Nottingham Trent University, United Kingdom

Corresponding Author:
Marijn Janssen, Delft University of Technology, Jaffalaan 5, Delft, Zuid-Holland 2628BX, the Netherlands.
Email: m.f.w.h.a.janssen@tudelft.nl

Introduction

The scale of use of artificial intelligence (AI) in decision making in the public and private sectors has been rapidly expanding in recent years (Duan et al., 2019). Research into AI in government is still scarce (de Sousa et al., 2019). Given the huge volumes of available data, AI algorithms and techniques can analyze and learn from the data in order to improve public services (Kankanhalli et al., 2019). The smartness created by AI can result in the creation of public values (Zhang et al., 2016). Much of the current research into AI is focused on the development and implementation of algorithmic systems for automated decision making. However, many of these AI applications have been criticized for their opaqueness and the lack of explanation of the causal link between input and output (Burrell, 2016). This lack of transparency hinders accountability and meaningful scrutiny of algorithmic solutions. Furthermore, it is hard to determine whether people affected by the use of algorithmic solutions are treated in a consistent manner and that no biases are introduced. In particular, society expects AI-based systems to adhere to public values like transparency, accountability, and equality. The above criticism has resulted in a resurgence of research into *explainable AI* (XAI; Miller, 2019), which refers to solutions based on algorithmic methods and techniques that can be understood by human experts. XAI embraces the social right to explain decisions to the public. The effects of AI and XAI on decision making have received scant attention.

Governments are becoming increasingly data-driven and are looking for ways to use these data in their decision-making processes. To this end, AI algorithms are being used more and more to deal with the variety and vast amount of data. The goal of the use of algorithms is often to augment decision making instead of achieving fully automated decision making. It has also been argued that algorithms can result in more rational decision making (de Sousa et al., 2019). The use of AI and algorithms has the potential to increase the speed of decision making and, at the same time, make better and more accurate decisions. Yet algorithms are not uniform and might yield different outcomes. The use of algorithms might not result in correct and accurate decisions (Barocas & Selbst, 2016). Algorithms can introduce inadvertent bias, reinforce historical discrimination, favor a particular political orientation, or reinforce undesired practices (Janssen & Kuk, 2016b). In this way, they could violate public values such as fair decision making and equal treatment of citizens without randomness or bias in decisions. The nature and structure of the algorithms themselves can inherently give rise to decisions based on statistically sound yet discriminatory, racist, or generally undesirable prejudices (Burrell, 2016). The use of any statistical data can result in false positives and false negatives. A false positive (or Type I mistake) is a result that indicates that a given condition is present when it is actually absent, whereas a false negative (or Type II mistake) improperly indicates the absence of a condition, when in reality it is present. For example, in the case of searching for a criminal, the search algorithm may seek to identify as many suspects as possible. The disadvantages are that persons will be unjustly classified as criminals, and more false positives will be created. As this is undesirable, the choice can be to decrease the number of false positives. This action can in turn, inadvertently, increase the number of false negatives and criminals might go undetected. Hence, trade-offs are needed about what is desirable from a societal point of view.

Algorithms increasingly make decisions that used to be made by people (Hong & Lee, 2018). There are many AI applications that support decision making in the public sector, ranging from those involved in handling a mundane task, such as the granting of a permit and calculation of tax returns, to those involved in performing more nuanced tasks, such as the granting of social benefits, the admittance of immigrants, and surveillance. The use of AI creates *data-driven government or algorithmic governance* (Janssen & Kuk, 2016a). And while there is a tendency to use algorithms for decision making more and more frequently, the influence of the use of algorithms on decision makers is less well understood. There is limited empirical research into the risks and benefits of AI

(Sun & Medaglia, 2019). The decision makers' role in the use of big data and algorithms is often overlooked (Arnaboldi, 2018). What's more, there seem to be high levels of overconfidence in computational solutions (Diesner, 2015).

The goal of the research presented in this article is to understand the effect of algorithms on human decision making. Governments from all over the world are embracing algorithms for automatic decision making. Specifically, we will look at decisions made by public servants in semiautomated decision making. This means that algorithms aid decision making while humans stay in control of making the final decisions. Humans' involvement is key when context-based factors need to be taken into account that cannot be found in the data when making decisions (Janssen & Kuk, 2016a). In this study, experienced and nonexperienced participants were asked to make decisions based on a given problem description and a recommendation formulated by an algorithm. In the experiment, the algorithm did not always provide the correct recommendation (there were false positives and false negatives), and it was subject to bias. The number of correct decisions was measured for three situations: without the use of any algorithm, with business rules (BR), and with machine learning (ML). In ML, a recommendation and a level of certainty are given. In BR, predefined rules are used to suggest a decision. In the Netherlands, several municipalities and the Dutch Immigration and Naturalization Service (IND) employ BR systems in their service provision processes. With the advent of AI, they are wondering if they should replace these systems by ML-based systems.

This article is structured as follows. In the next section, the literature background is presented, followed by the formulation of hypotheses. The design of the experiment is presented in the "Research Approach" section. Thereafter, the findings are presented, followed by a discussion of the findings. Finally, conclusions are drawn.

Algorithmic Decision Making

There is a long tradition of literature focusing on decision making in public administration and government. The wicked nature of the problems faced by public administrators or governments is often emphasized: the fact that they often need to be solved based on incomplete, contradictory, and changing information (Churchman, 1967; Rittel & Webber, 1973). Simon (1972) proposed the theory of bounded rationality, which results from the intractability of natural decision-making problems, the cognitive limitations of the mind of human decision makers, and the limited time available to make the decision. Lindblom (1959) suggested that decision making is not limited by bounded rationality, but that no theories are used at all and that decisions are made quickly based on very little information.

With the rise of the Internet, citizens increasingly expect services to be provided online and decisions about granting or rejecting applications, for example, to be made within a short time frame. This has resulted in the automation of decision making and the use of algorithms for making decisions. Our work focuses on semiautonomous decision making with the use of algorithms, which is often embraced for situations in which a large number of decisions need to be made on a large scale within a very short time. Algorithms are being used for decision making more and more. A key premise is that better information will lead to better decision making (Höchtl et al., 2016 and in the subsequent occurrences.]). This embraces the ideal view of bureaucratic decision making as highly organized and thorough and involves systematic approaches to making decisions. Often terms like evidence-based decision making and data-driven decisions are used to refer to this field (Zhang et al., 2016). These terms suggest that this kind of decision making is more rational and that data are facts. Reality is more cumbersome, though, and algorithms and AI are much criticized for their susceptibility to bias and inability to deal with variety (Janssen & Kuk, 2016b; Selbst et al., 2019).

Much current research is focused on what is often called “narrow artificial intelligence” (Goertzel & Pennachin, 2007). This type of AI is goal-focused and typically performs a single task (Goertzel & Pennachin, 2007). Algorithms that learn by identifying patterns in data sets are often used. ML that uses data to infer and overtake structured human tasks receives a lot of attention today (Brynjolfsson & Mitchell, 2017). Most approaches use ML by looking at the inputs and outputs (Selbst et al., 2019). ML algorithms are often criticized for their opacity (Burrell, 2016). Selbst et al. (2019) challenged the notions of fairness and discrimination in the societal context that surrounds ML-based decision making. The opacity makes it hard to scrutinize decisions and to ensure that decisions are always made by applying the same logic. Furthermore, ML heavily depends on the quality of the data. Data can be biased due to various institutional practices, the background, and culture of the data subjects whose data are gathered. In this way, inadvertent bias can be introduced (e.g., Kashin et al., 2015), and historical discrimination reinforced (Kroll, 2015), or outdated practices or failures from the past might be reinforced while societal values and political preferences have changed.

Another stream within the AI-based decision-making field is the use of expert systems with BR, which originates from expert systems that gained popularity in the 90s (Duan et al., 2019; Hurley & Wallace, 1986). Expert systems aimed to replicate human behavior by defining declarative logic (Turban & Watkins, 1986). This logic is used to represent human decision making in the public sector (Hurley & Wallace, 1986). The BR should reflect and codify legislation. Whereas current AI applications are often data-driven, expert systems are more knowledge-driven. These systems often represent knowledge concerning: (1) facts about the domain, (2) hard rules or procedures, (3) problem situations and potential solutions, (4) general strategies, and (5) conceptual models of the domain (Hurley & Wallace, 1986).

Although ideally, legislation should be reflected in BR in a one-to-one manner, reality is more cumbersome. Usually, legislation is written at a high level of abstraction, and interpretations are needed to translate them into BRs (Gong & Janssen, 2013). There often are many exceptions that might not be easily captured by BR. Furthermore, the large number of BRs adds to the complexity, which complicates maintenance and understanding. In the Netherlands, many municipalities and several administrative organizations make use of BR in their decision-making processes.

Whereas BR was used in the past, ML has claimed much attention for automatic decision making recently. ML and BR have similar intentions, and both are used nowadays to support decision making, but they take different approaches, as shown in Table 1.

ML comprises computational algorithms that are able to learn from data without having to be given explicit instructions. ML is a general-purpose technology that is suitable for various tasks (Brynjolfsson & Mitchell, 2017). It is typically classified into supervised, semisupervised, and unsupervised learning (Jordan & Mitchell, 2015). In supervised learning, the input *and* output data are available, whereas in unsupervised learning, only the input data are available. ML uses training data to infer relationships between data. Jordan and Mitchell (2015) found that for many applications, it is easier to train a system by showing it examples of desired input–output behavior than to program it manually. The latter requires analyzing the desired input and the corresponding responses and might be time-intensive. ML is often criticized for not explaining how it arrives at a certain prediction.

Most ML algorithms focus on function approximation problems (Jordan & Mitchell, 2015). The tasks are embodied in an input/output function like identifying whether somebody should be granted a permit or not. ML can help to improve the accuracy of such a function. The question is what the accuracy of the output is and how it can be used for decision making.

Of late, there has been a lot of focus on *XAI*, which explicitly explains decisions or actions to a human observer (Miller, 2019). The need for explanation originates from the bias, mistakes, and lack of trust in AI systems (Butler, 2013). XAI refers to “an explanatory agent revealing underlying causes to its or another agent’s decision making” (Miller, 2019, p. 4). Miller differentiated between

Table 1. Summarizing the Differences Between Machine Learning and Business Rules.

	Machine Learning	Business Rules
Creation of knowledge base	Inference of patterns from data (without explicit instructions)	Defining explicit instruction by human beings based on logic and assumed causal relationships
Implementation	Learning from historical data	Formalization of rules
Input	Historical data	Law and rules + case documents + experience + interpretation
Typical risks	– Inheriting bias from historical data – Unable to deal with exceptions and unusual cases – Dependent on the quality of learning data – Limited accuracy in outcomes	– Interpreting legislation differently than intended – Unable to deal with noncoded situations – High complexity of rules which are not easy to understand, to analyze conflicts and irregularity and to adapt
Explainability	Opacity	Declarative logic

interpretability or explainability and explanation. Interpretability or explainability refers to how well a human could understand the decisions in the given context, and explanation refers to explicitly explaining decisions to people. In our situation, we look at the explainability, as the algorithms should suggest a decision and explain why it was taken.

BRs define the decision-making logic by defining declarative logic. Rules reduce the need for programming, that is, the amount of code that needs to be written, making it easier to read and change (Von Halle, 2001). The rules explain how the system arrives at a decision; the decision maker can have complete knowledge of the causal rules. BR approaches have become a common approach for use in decision making (Vanthienen, 2001). The *Semantics of Business Vocabulary and Business Rules* (SBVR) is a standard developed by the Object Management Group. SBVR can be used for declarative descriptions in the natural language. It is used to articulate the rules for making a decision. The rules are highly dependent on the available information and on the context. If information is not correct, the decision might be wrong. If not all information that is needed is available, it will be harder to arrive at a sound decision. Furthermore, some rules might not be appropriate in a certain context or might give the wrong outcomes, as they might be context-dependent.

Hypothesis Formulation

Algorithms are used to automate decision making. Although fully automated decision making is possible, we focus on situations in which algorithms are used to support decision makers, that is, situations in which human beings make the final decision. We expected that the use of algorithms for supporting the decision-making process would result in more correct decisions (Höchtl et al., 2016). Therefore, we formulated the following hypothesis.

Hypothesis 1: The use of algorithms will result in more correct decisions.

There is a range of algorithms in use that are based on supervised, semisupervised, unsupervised, and reinforcement learning (Alpaydin, 2009; Russell & Norvig, 2016). Algorithms differ in their level of explainability (Adadi & Berrada, 2018). We also expected that the type of algorithm would matter in supporting decisions. Algorithms vary in their effectivity, and different algorithms may yield different results (Russell & Norvig, 2016). Both BR and ML have their merits and shortcomings as listed in Table 1. They are developed in different ways. This brings us to our second hypothesis.

Hypothesis 2: The type of algorithm influences the outcome of decision making in the same cases.

Algorithms are criticized for being too opaque in their decision making (Brauneis & Goodman, 2018). Explanation is needed to facilitate understanding and learning (Miller, 2019). When the recommendations are not transparent, decision makers might find it more difficult to understand the logic and reasoning and might not be able to scrutinize the recommended decision. Hence, opaque algorithms will result in more mistakes. The BR approach enables decision makers to understand the logic of decision making and track back why a certain decision was made. This is used for measuring a *transparent* algorithm. The ML approach is a black-box approach in which decisions are guided by estimates of the ML outcome's uncertainty. The ML approach represents an algorithm that is opaque, for example, less transparent. We expect that ML will be less effective because of its opaqueness than BR in helping to detect mistakes. Therefore, the following hypothesis was formulated.

Hypothesis 3: The use of transparent algorithms results in more correct decisions.

Raghunathan (1999) found that the quality of decisions improves if the decision maker understands the problem and the relationships between problem variables. Besides, decision makers' skills and their experience proved important in decision making (Dwivedi et al., 2017). In order to make a decision, the decision makers should understand the limitations and the potential for their decision making. As such, the experience and understanding of the decision makers might affect the outcomes. We will take this variable into account in our research approach by measuring familiarity with the domain. Those who are already familiar with the domain will make better decisions than those without any prior experience. We hypothesized that experienced persons are better at detecting wrong decisions made by algorithms.

Hypothesis 4: Experienced persons will make more correct decisions than nonexperienced persons.

These four hypotheses will be tested by conducting an experiment, which will be discussed in the next section.

Research Approach

In this study, we used a quasi-experiment to evaluate the effect of algorithms' use for decision making. The study's design compared human decision making with two other conditions—BR and ML. In effect, the human decision-making condition served as the comparison group. Besides, we also compared decision making by experienced persons with that by nonexperienced persons.

For the experiment, we aimed to select an organization that was already using automated decision making. In the Netherlands, the Dutch IND has been experimenting with automated decision making for a long time and a lot of information about their decision making is publicly available on their website (https://ind.nl/). Their automated decision-making system went into operation in 2010. The asylum request and residence permit application for the Netherlands were selected because these decisions are highly sensitive and not straightforward and often require human involvement in decision making. Public information is available like the evaluation of procedures and backgrounds (Böcker et al., 2014) and architecture descriptions (Gong & Janssen, 2012). Use cases were derived and descriptions were included as close as realistically possible. An overview of the attributes influencing the decisions was made, which includes the land of origin, age, gender, profession, income, and specifics such as motivation for coming to the Netherlands, religion (if applicable) and political convictions, and any other reasons for the asylum request or visa application. The rules used by the IND were analyzed and provided as an input for both types of algorithms. We only used those cases in which the decisions were not contested to ensure that a correct decision was possible.

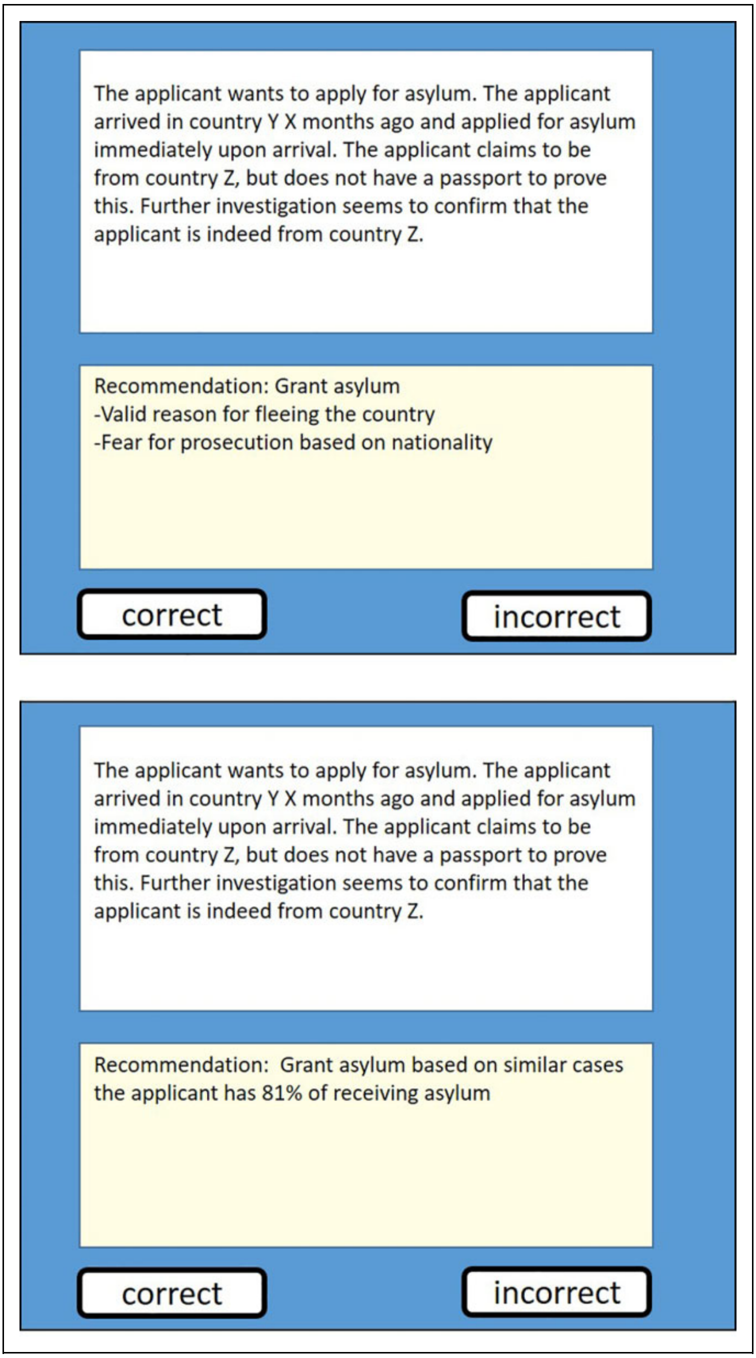
From the use cases, we derived BRs and used supervised ML to learn to make decisions to grant or reject an asylum request or residence permit application. We introduced exceptions and ambiguity in the use-case descriptions. Consequently, BRs were not able to arrive at a correct decision for all situations, as not all exceptions were codified in BRs and the BRs were not able to interpret the ambiguity in some of the formulations. When predicting a decision, ML sometimes made the wrong decisions due to mistakes in the learning data and its inability to discriminate between subtle differences between the use cases. Besides, some more mistakes, ranging from obvious to complex ones, were introduced by providing the wrong input data. Furthermore, some BRs were used that were not appropriate in the context and would subsequently give the wrong outcome. Finally, false positives and negatives were included in the experiment. This resulted in BR and ML recommendations for making a decision being incorrect.

If a permit was granted or denied in accordance with legislation by the decision maker, then the decision was marked as "correct." If a permit was granted when it should not have been granted, or if a permit was denied when it should have been granted, then the decision was marked as "incorrect." Figure 1 shows an example of the interface with BR and ML. The text in this figure is translated from the original text in Dutch. Whereas BR provided a single suggestion in our experiment, ML provided a suggestion and an estimate of the accuracy of the recommendations made by ML.

In the experiment, algorithms were trained using valid and incorrect (skewed) data to ensure that both correct and incorrect results would be included. Some context-based factors (which were not included in the algorithm but are of importance to the decision making) were included, which could result in a wrong recommendation for a decision by the algorithm. The participants were asked to make a decision for each case as to granting or refusing requests for asylum or granting or rejecting a residence permit application. Before they started, the participants received instructions, and the applicable legislation was presented to enable the participants to make the correct decisions. Simple and more difficult cases were presented in random order to avoid learning effects biased toward a certain outcome. The participants were given 30 different cases. In the situation of human decision making without any algorithm, only information was given and an explanation of how decisions should be made. With BR, a recommendation for a decision was presented. The participants could look at the rules used for arriving at a decision and the input data for making these decisions. With ML, a recommendation for a decision was made as well as the certainty of the decision. The participants could look at the case at hand to investigate manually if a decision was correct. Based on this input, the participants were asked to make the correct decisions. For this, they needed to judge whether the decision made by BR or ML would be correct or wrong.

The experiment was conducted online with all the information being shown on a computer; meanwhile, the participants were invited and instructed by a student. The participants had to decide on an immigration request. The participants could either grant or reject the applications; no alternative options were given, with the result that the participants were forced to decide for each case.

The participants were instructed to make the best decisions. Making all the decisions took between 20 and 75 min. Several participants did not finish the experiment, as it took them too long. We continued until we had 30 experienced and nonexperienced participants in total. Finding sufficient public decision makers to arrive at statistically significant outcomes was found not to be feasible at the Dutch IND; instead, we asked students enrolled in master's programs that were oriented toward working for the government. We conducted the experiments with a few civil



The applicant wants to apply for asylum. The applicant arrived in country Y X months ago and applied for asylum immediately upon arrival. The applicant claims to be from country Z, but does not have a passport to prove this. Further investigation seems to confirm that the applicant is indeed from country Z.

Recommendation: Grant asylum
-Valid reason for fleeing the country
-Fear for prosecution based on nationality

The applicant wants to apply for asylum. The applicant arrived in country Y X months ago and applied for asylum immediately upon arrival. The applicant claims to be from country Z, but does not have a passport to prove this. Further investigation seems to confirm that the applicant is indeed from country Z.

Recommendation: Grant asylum based on similar cases the applicant has 81% of receiving asylum

Figure 1. Examples of the business rules (top) and machine learning (bottom) interfaces.

servants from a large administrative organization. When comparing the results, we found no reason for these results being different, but this cannot be statistically proven. The master program includes various aspects, such as policy making, organization and management, statistics, and evidence-based decision making. Most of the students were between 22 and 30 years old. Only four students

Table 2. Setup of the Experiment.

	Human-Based Decision Making	ML-Supported Decision Making	BR-Supported Decision Making	Total
Nonexperienced participants	30	30	30	90
Experienced participants	30	30	30	90
Total	60	60	60	180

Table 3. Means of the Numbers of Correct Decisions.

	Human-Based Decision Making	ML-Supported Decision Making	BR-Supported Decision Making	Total
Nonexperienced participants	21.20	23.13	24.33	22.89
Experienced participants	25.20	26.17	27.27	26.21
Total	23.20	24.65	25.80	24.55

were older than 30 years. About one quarter of the students claimed to have prior experience in public administration. The international students were categorized as experienced students, as these students had the experience of applying for a visa when coming to the Netherlands.

Each time, the participants would be given a case either supported by no means, ML, and BR. Only the advice given by the algorithm was presented, and some background about the working of the algorithms was given at the beginning. The sequence of decisions to be made was randomized. No feedback was given during the session about whether decisions were made correctly or wrongly. Half of the suggestions presented by the algorithm were wrong. The number of wrong decisions was the same for ML and BR. The setup of the experiment is shown in Table 2. Analysis of variance (ANOVA) was used to analyze whether the use of different algorithms resulted in differences in decisions and whether there were differences among decisions made by experienced and nonexperienced participants.

Findings

An ANOVA was conducted to evaluate the experiments. In the following tables, the means, variation, and ANOVA will be discussed. Table 3 shows the means of the correct number of answers from the experiments. Most failures were made when only humans made the decisions without any algorithmic support, as in total, there were only 23.20 on average correct out of the 30 in total. The average shows that both ML and BR result in more correct decisions made by the participants. The least mistakes were found when using BR. One reason for this could be that when using BR the reasoning for making decisions is transparent, and it is easier to detect mistakes. Not surprisingly, more experienced persons made fewer mistakes and made better decisions than those without any experience. We did not look at learning over time, for example, whether people arrived at better decisions after having made more decisions during the experiment. Some participants suggested that they learned during the experiment. This is an interesting avenue for further research to determine whether certain algorithms help arrive at better decisions.

Table 4 shows the variance of the number of correct decisions. The variance is the biggest in the group of human decision making without any experience. This suggests that the making of correct decisions was not uniformly distributed and had a high spread for human-based decision making. With the use of algorithms, the variance decreases, as shown in Table 4. This suggests that

Table 4. Variance of Correct Decisions Made by the Participants.

	Human-Based Decision Making	ML-Supported Decision Making	BR-Supported Decision Making	Total
Nonexperienced participants	14.92	7.43	6.37	11.04
Experienced participants	8.72	6.21	6.34	11.67
Total	15.69	9.04	8.30	12.03

Table 5. ANOVA Overview.

ANOVA				α	0.05	
	SS	*df*	*MS*	*F*	*p Value*	*Sig.*
Rows	496.67	1	496.67	421.42	0.002	yes
Columns	203.70	2	101.85	86.42	0.011	yes
Inter	2.36	2	1.18	0.14	0.868	no
Within	1,449.82	174	8.33			
Total	2,152.55	179	12.03			

Note. ANOVA = analysis of variance.

algorithmic support helps the participants make more consistent decisions. BR has the lowest variance, which suggests that BRs are most suitable for ensuring the making of consistent decisions.

Table 5 shows the results of the ANOVA. The *F* value is high, suggesting that the samples are independent. The *p* values show that the differences between the human, ML, and BR-based decision making are significant. Also, the difference between experienced and nonexperienced persons is found to be significant.

The findings show that understanding of decisions combined with experience is the most important factor for humans being able to make correct decisions supported by algorithmic decision making. However, even experienced persons were not able to identify all mistakes. Several BR and ML cases that were incorrect were not identified by the participants and assumed to be correct. There was no single case that was dominated by incorrect decisions. Our first hypothesis, "The use of algorithms will result in more correct decisions," is confirmed, as both ML and BR resulted in more correct decisions when compared with human-based decision making.

A difference was found in the types of algorithms supporting the decisions as BR resulted in more correct decisions than ML in our experiment. We also expected that the type of algorithms would matter in supporting decisions. Our second hypothesis was confirmed, as there was found to be a significant difference in the average number of correct decisions between ML and BR, suggesting that "The type of algorithm used influences the outcome of decision-making in the same cases." This confirms that algorithms vary in their effectivity, and different algorithms yield different results. Public organizations should be aware of which type of algorithm they use and should not blindly select one.

Whereas ML is characterized by its opacity, the BR approach enables decision makers to understand the logic of decision making and to track back the reasons why a certain decision is made. The ML approach is a black box approach in which decisions are guided by estimates of the uncertainty of the ML outcome. The ML approach represents an algorithm that is less transparent than the BR approach. With the BR approach, more correct decisions were made than with the ML approach, which supports our third hypothesis that "The use of transparent algorithms results in more correct decisions." This suggests that the use of understandable and transparent algorithms by governments is important to ensure that decisions are made correctly.

Finally, we expected that experience with the type of decision that needed to be made would play an important role. Decision makers who are already familiar with the domain will make better decisions than those without any prior experience (Raghunathan, 1999). Those with experience were considered to be those participants who had applied for a residence permit for the Netherlands. We hypothesized that "experienced persons will make more correct decisions than non-experienced persons." This was confirmed, as experienced persons were better at detecting wrong decisions made by algorithms.

Discussion

AI algorithms are used more and more by governments to support or even make decisions. The goal of the research presented in this article is to understand the effects of the use of algorithms on human decision making. Table 3 shows that the use of algorithms helps decision makers to make more correct decisions, whereas Table 4 suggests that algorithmic support helped the participants to make more consistent decisions. These findings confirm the reasons for introducing AI in decision making to improve decision quality (Zhang et al., 2016). However, the findings also show that identifying incorrect recommendations made by algorithms is challenging and that algorithms might easily blind people. The participants were not able to detect all wrong suggestions made by the algorithm.

Algorithms are criticized for being too opaque for supporting decisions (Brauneis & Goodman, 2018), which would make it hard for decision makers to interpret their outcomes. A difference was found between judgments with the aid of BR or ML. This confirms that algorithms vary in their effectivity, and different algorithms might yield different results (Russell & Norvig, 2016). People identified more mistakes in BR than in ML, while the number of mistakes was equal. It appears that the explainable nature of BR can explain this phenomenon. This finding suggests that, both for white-box (that can account) and black-box AI algorithms, decision-making mistakes can be made. However, with a white-box approach (BR), more mistakes were found by the participants, as in total more correct decisions are made. Yet, ensuring the ability to understand and traceback decisions is not sufficient to avoid all mistakes and make only correct decisions. There were only a few participants who made no or hardly any mistakes. These were primarily found in the experienced group, which suggests that experienced decision makers make more correct decisions. This confirms the findings of Raghunathan (1999) that experience influences decision maker quality, although decision maker quality involves many other aspects that were not covered in our research.

In the literature, it is unclear what meaningful "algorithmic transparency" entails (Brauneis & Goodman, 2018). There are various ways of defining this. In our work, we measured it as the number of incorrect decisions made by the participants. When using algorithms, the decision makers followed the wrong recommendations. Despite the experiment's relatively simple nature, only a small portion of the participants made all decisions correctly in the experiment. This suggests that XAI will not result in transparency per se. The complexity and the wicked nature of the problems faced might simply prevent the creation of transparency. Furthermore, there might be variables like the context and mistakes in the input data that is hard to detect and to take into account.

Learning from and updating the data to generate better insights was outside the scope of this research. For this, Bayes's rules in business analytics can be used to update the prior distribution to calculate the posterior distribution of predictive models (Janssen & Kuk, 2016a). Instead of having general models, the probability of an event occurring is based on conditions. In this way, the models can become more accurate over time. This requires the involvement of humans to provide feedback and update the models. This is an important research value for improving the working of algorithms.

In our experiment, the input data were predefined and structured, and we only selected those cases in which a clear decision could be made. We excluded situations in which the decision could be contested. The veracity and variety of data are not taken into account, which adds additional complexity and challenges for interpretation and the sense-making of the data. Sense-making is the process of giving meaning to an ambiguous situation (Weick et al., 2005). We recommend further researching the processes that constitute the meaning of the decision and the enacted behavior, given an ambiguous situation that might have diverse data as input.

Algorithms are not empathetic. Data fed into the algorithms for learning determine its outcome and ultimately the quality of the data affect the decisions (Janssen et al., 2020). Any algorithm will eventually have biases, as it is based on historical data to make predictions about the future, whereas, in the meantime, the situation might have changed. We purposely introduced incorrect suggestions and mistakes in the algorithms to determine whether the suggestions were followed and whether decision makers would spot these mistakes. Different types of algorithms call for different types of presentation and support to prevent bias and ensure sound decision making. In our experiment, both were presented in simplified ways. An advanced user interface and more information can help to interpret these algorithms better. For example, the system might draw the attention to possible mistakes or shortcomings of the data and algorithm. Although we did not look into different types of ML and BR, it is likely that these differences also matter. Further research can take a more fine-grained look at algorithms.

The experiment can be criticized for being too simplistic and as being a setup for making wrong decisions. The participants had to make a decision, and the algorithms often recommended a wrong decision, whereas the participants had no information in advance about the number of times the algorithms would be correct. Although such starting points might not hold in practice, they give a good indication of the possible effects when there is time pressure. Furthermore, the experiment was fairly simple and used only two types of algorithms. The participants had to select a decision and could not opt to investigate further when in doubt. We designed it in this way to ensure that the participants had to make a choice. Some of the participants did not like this and quit during the experiment, whereas some others commented on this as being annoying after the session ended. More advanced ways could be created; however, this would require the involvement of a larger sample, whereas we already struggled to get sufficient participants for the experiment.

The participants often found it difficult to make decisions. The variance in time when making decisions was bigger than initially expected. Some participants even took four times longer to make decisions than other participants. Although we did not measure the time, our observations suggest that experienced persons' decision-making speed is higher than that of nonexperience. Furthermore, the decision-making time seemed to differ per algorithm. We recommend further research into the speed of decision making relative to the quality of the decisions.

From the findings, implications for policy makers and decision makers can be drawn. Although some argue that AI algorithms might outperform humans (Grace et al., 2018), algorithms are not flawless. Furthermore, in AI-supported situations, even humans might not be able to detect all mistakes. This suggests that probably not all decision-making processes should be supported by algorithms. Only those that are suitable for automated decisions can be supported, and exceptions can be dealt with manually. As no single decision maker was able to detect all mistakes, we suggest having more than one person to be involved in this decision making to ensure that mistakes are detected and in this way, raise the quality of decisions.

Another implication of our research is that data, the decision-making process design, and algorithmic governance are key elements to ensure proper outcome (Janssen et al., 2020). Sound data are the basis for ML algorithms, and ensuring accurate and unbiased data is essential. ML is used in a predictive sense, whereas input data are based on prior decisions that might not be correct or biased, making it hard to evaluate counterfactual decision (Kleinberg et al., 2018). Preventive and detective controls should be added to the whole process of using algorithms to ensure the finding of mistakes, bias, and other flaws (Janssen et al., 2020). A key element is how the use of algorithms is designed and how people deal with this.

Furthermore, the findings imply that selecting the appropriate algorithms for supporting decisions is crucial. Whereas the current hype is for embracing ML for making decisions, there are other strands in AI that might provide more appropriate algorithms. Those who employ BR systems should be reluctant to move to ML systems, as the data-driven approach might not be better or even provide worse outcomes than the BR approach. Specifically, our experiment shows that the research and experiences from the BR strand can be used to create a higher level of fairness, transparency, and accountability. Instead of simply jumping on the bandwagon, algorithms should be chosen by analyzing the problem at hand and understanding the public values that apply. Selecting inappropriate algorithms might have far-reaching consequences.

Probably, the combination of different computational algorithms provides better results than a single computational algorithm. The use of a certain algorithm is dependent on the objective, the desired outcomes, and the context. Finally, once a system is in place, training of decision makers is key to their ability to find mistakes and thus increasing accountability and transparency.

Conclusions

The goal of our research is to understand better the effect of algorithms on human decision making by public servants in government. We investigated this by conducting an experiment in which the participants had to make decisions guided by algorithms. This enabled us to control other factors and look at the effects of different algorithms in recommending a decision and investigating decision makers' experiences at the same time. Participants could make correct or wrong decisions. Bias and mistakes were introduced in the algorithms to fool the participants. The findings show that understanding of decisions combined with experience is the most important factor in humans being able to make correct decisions supported by AI algorithms. All too often, the assumption is that if algorithms are well designed, transparency and accountability are created. In this article, it is demonstrated that the type and inherent characteristics of algorithms matter. Our experiment suggests that governments should not jump on the bandwagon and move from rule based to data-driven computational algorithms, as this might reduce explainability. Furthermore, experienced persons made better decisions, which suggests that experience and education are important when working with algorithms for decision making. Finally, we found that explainability will not result in transparency and better decision making per se, but it contributed to more accurate decisions being made in our experiment. Hence, there is a risk that algorithms might blind people and incorrect decisions will still be made.

Our work suggests various avenues for further research. An initial suggestion is that the research can be generalized for different types of decisions and algorithms. Different algorithms have different impacts on decisions made by human decision makers. In particular, classification of the type of decisions and classification of the type of algorithms can be used to gain deeper insight into the variations and different effects of algorithms. The complexity of the decision-making context might influence the suitability of the use of algorithms. More research into the effect of data quality, the ambiguity of legislation, and other factors on algorithms is needed.

Our focus was on investigating the effect of algorithms on the correctness of decisions, and we did not look at other factors such as trust, fairness, transparency, or accountability of algorithms. These are complex constructs, and their effects on algorithmic decision making should be further researched. We did not investigate the effect of the user interface on decision making. However, a good user interface can prevent, especially nonexperienced, decision makers from making mistakes. More advanced user interfaces might help interpret the results better and might draw attention to detect typical mistakes. Finally, pressure on the amount of time provided to decision makers might affect the quality of decisions, and we recommend researching its effect. This can all help to improve the quality of AI-supported algorithmic decision making.

In our experiment, no single decision maker was able to detect all mistakes made by algorithms. More transparent algorithms would result in better detection of mistakes. This suggests that the use of algorithms should be done with care, and governments should be prudent to use ML types of algorithms for decision making. BR might better address explainability and other public sector requirements than data-driven computational algorithms do. Which type of algorithm is suitable for which problem should be further researched?

Authors' Note

The authors want to acknowledge the students who have helped to develop the prototypes and to collect the data.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

References

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, *6*, 52138–52160.

Alpaydin, E. (2009). *Introduction to machine learning*. MIT Press.

Arnaboldi, M. (2018). The missing variable in big data for social sciences: The decision-maker. *Sustainability*, *10*(10), 3415.

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review*, *104*, 671.

Böcker, A., Grütters, C., Laemers, M., Strik, M., Terlouw, A., & Zwaan, K. (2014). *Evaluatie van de herziene asielprocedure: Eindrapport* [Evaluation of the revised asylum procedure: Final report]. Radboud University.

Brauneis, R., & Goodman, E. P. (2018). Algorithmic transparency for the smart city. *Yale Journal of Law & Technology*, *20*, 103.

Brynjolfsson, E., & Mitchell, T. (2017). What can machine learning do? Workforce implications. *Science*, *358*(6370), 1530–1534.

Burrell, J. (2016). How the machine "thinks": Understanding opacity in machine learning algorithms. *Big Data & Society*, *3*(1). https://doi.org/10.1177/2053951715622512

Butler, D. (2013). When Google got flu wrong. *Nature*, *494*(7436), 155.

Churchman, C. W. (1967). Free for all. *Management Science*, *14*(4), 141–146.

de Sousa, W. G., de Melo, E. R. P., 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), 101392. https://doi.org/10.1016/j.giq.2019.07.004

Diesner, J. (2015). Small decisions with big impact on data analytics. *Big Data & Society*, *2*(2), https://doi.org/10.1177/2053951715617185.

Duan, Y., Edwards, J. S., & Dwivedi, Y. K. (2019). Artificial intelligence for decision making in the era of Big Data–evolution, challenges and research agenda. *International Journal of Information Management*, *48*, 63–71.

Dwivedi, Y. K., Rana, N. P., Janssen, M., Lal, B., Williams, M. D., & Clement, M. (2017). An empirical validation of a unified model of electronic government adoption (UMEGA). *Government Information Quarterly*, *34*(2), 211–230. https://doi.org/10.1016/j.giq.2017.03.001

Goertzel, B., & Pennachin, C. (2007). *Artificial general intelligence* (Vol. *2*). Springer.

Gong, Y., & Janssen, M. (2012). From policy implementation to business process management: Principles for creating flexibility and agility. *Government Information Quarterly*, *29*, S61–S71.

Gong, Y., & Janssen, M. (2013). An interoperable architecture and principles for implementing strategy and policy in operational processes. *Computers in Industry*, *64*(8), 912–924.

Grace, K., Salvatier, J., Dafoe, A., Zhang, B., & Evans, O. (2018). When will AI exceed human performance? Evidence from AI experts. *Journal of Artificial Intelligence Research*, *62*, 729–754.

Höchtl, J., Parycek, P., & Schöllhammer, R. (2016). Big data in the policy cycle: Policy decision making in the digital era. *Journal of Organizational Computing and Electronic Commerce*, *26*(1–2), 147–169. https://doi.org/10.1080/10919392.2015.1125187

Hong, S., & Lee, S. (2018). Adaptive governance, status quo bias, and political competition: Why the sharing economy is welcome in some cities but not in others. *Government Information Quarterly*, *35*(2), 283–290.

Hurley, M. W., & Wallace, W. A. (1986). Expert systems as decision aids for public managers: An assessment of the technology and prototyping as a design strategy. *Public Administration Review*, *46*, 563–571.

Janssen, M., Brous, P., Estevez, E., Barbosa, L. S., & Janowski, T. (2020). Data governance: Organizing data for trustworthy artificial intelligence. *Government Information Quarterly*, *37*(4), 1–8. https://doi.org/10.1016/j.giq.2020.101493

Janssen, M., & Kuk, G. (2016a). Big and Open Linked Data (BOLD) in research, policy and practice. *Journal of Organizational Computing and Electronic Commerce*, *26*(1–2), 3–13. https://doi.org/10.1080/10919392.2015.1124005

Janssen, M., & Kuk, G. (2016b). The challenges and limits of big data algorithms in technocratic governance. *Government Information Quarterly*, *33*(3), 371–377. https://doi.org/10.1016/j.giq.2016.08.011

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

Kashin, K., King, G., & Soneji, S. (2015). Explaining systematic bias and nontransparencyin US social security administration forecasts. *Political Analysis*, *23*(3), 336–362. https://doi.org/10.1093/pan/mpv011

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2018). Human decisions and machine predictions. *The Quarterly Journal of Economics*, *133*(1), 237–293.

Kroll, J. A. (2015). *Accountable algorithms*. Doctoral Dissertation Princeton University. http://dataspace.princeton.edu/jspui/handle/88435/dsp014b29b837r

Lindblom, C. E. (1959). The science of "muddling through." *Public Administration Review*, *19*, 79–88.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*, 1–38. https://doi.org/10.1016/j.artint.2018.07.007

Raghunathan, S. (1999). Impact of information quality and decision-maker quality on decision quality: A theoretical model and simulation analysis. *Decision Support Systems*, *26*(4), 275–286. https://doi.org/10.1016/S0167-9236(99)00060-3

Rittel, H. W. J., & Webber, M. M. (1973). Dilemmas in a general theory of planning. *Policy Sciences*, *4*(2), 155–169. https://doi.org/10.1007/BF01405730

Russell, S. J., & Norvig, P. (2016). *Artificial intelligence: A modern approach*. Pearson Education Limited.

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

Simon, H. A. (1972). Theories of bounded rationality. *Decision and Organization*, *1*(1), 161–176.

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

Turban, E., & Watkins, P. R. (1986). Integrating expert systems and decision support systems. *MIS Quarterly*, *10/2*, 121–136.

Vanthienen, J. (2001). Ruling the business: About business rules and decision tables. *New Directions in Software Engineering*, 103–120. https://www.semanticscholar.org/paper/Ruling-the-Business-%3A-About-Business-Rules-and-Vanthienen/8aaee7079dd06b75ad60e3a8535e5cf23752f1a3

Von Halle, B. (2001). *Business rules applied: Building better systems using the business rules approach*. Wiley Publishing.

Weick, K. E., Sutcliffe, K. M., & Obstfeld, D. (2005). Organizing and the process of sensemaking. *Organization Science*, *16*(4), 409–421.

Zhang, P., Zhao, K., & Kumar, R. L. (2016). Impact of IT governance and IT capability on firm performance. *Information Systems Management*, *33*(4), 357–373.

Author Biographies

Marijn Janssen is a full Professor in ICT & Governance at the Information and Communication Technology section of the Technology, Policy and Management Faculty of Delft University of Technology. More information: www.tbm.tudelft.nl/marijnj.

Martijn Hartog is a PhD candidate at the Faculty of Technology, Policy & Management at Delft University of Technology and a project- and programme manager for the Dutch government.

Ricardo Matheus is a lecturer and researcher in the field of Open government Data and Infrastructures at the Information and Communication Technology research group of the Technology, Policy and Management Faculty of Delft University of Technology.

Aaron Yi Ding is leading the Cyber-Physical Intelligence (CPI) Lab in the Department of Engineering Systems and Services at TU Delft. He is a tenure-track assistant professor at TU Delft and adjunct professor (permanent) in Computer Science at University of Helsinki.

George Kuk is a full professor of innovation and entrepreneurship at the Nottingham Business School.