Social acceptance of machine learning in a decision making context

Decision-making in a Complex Adaptive System Setting
(5294DMIA6Y)

Marit Beerepoot, 10983430
Jessy Bosman, 11056045

Abstract

With the amount of data growing each day, the use of algorithms seems to shift from a supportive role of data processing to a decision making tool. In the past research has been done to identify the dangers that machine learning in decision making might cause. A popular critique is that it is an error-prone and opaque process, that can be biased by technical, social and emergent biases. What a lot of people seem to forget, is that even though human decision making is socially accepted, it can also be biased and is not necessarily the better option. This research explores the field of social acceptance of machine learning in a decision making context, since this was not explored before. It is researched whether social acceptance is situationally dependent, if people are aware of biases and if there is a correlation between social acceptance and perceived biases. It was found that the majority seems to be positive about the general use of machine learning and is not afraid of it. There exists a general trust in and acceptance of machine learning. The acceptance of machine learning, however, seems to be situation dependent. Next to that, people seem to be aware of biases introduced by machine learning, especially the technical biases. Furthermore there does not seem to be a relation between the acceptance and the biases. Finally, preferably a combination of both machine learning and human cognition are used for decision making.

Keywords

Machine learning – Human biases – Algorithmic biases – Decision making – Social acceptance

1. Introduction

The amount of data that is generated daily is growing more rapidly than ever. Currently, 2.5 quintillion bytes of data are created every day (Marr, 2018). It is not possible to make sense of all this data without the help of algorithms to gain (new) insights (Flach, 2012). With the increasing amount of data, a shift in the use of algorithms is visible: algorithms used to be a supportive tool but are now more and more used as a decision making tool (Danaher et al., 2017; Zarsky, 2016). Even though the use of algorithms as a decision making tool is new, speculations about future applications seem endless. Domingos (2015) for example dreams of a 'master algorithm' that can learn and adapt to every decision-making situation without the help of human control.

In the past a lot of research done into the policy concerns and algorithmic governance that are emerging from using algorithms to make decisions (Danaher et al., 2017; Zarsky, 2016). There is also research done into the concerns and possible dangers of an algorithmic decision making process (Mittelstadt et al., 2016). Especially the opacity and transparency of the algorithms is a popular research point that is also often used as an argument to not implement algorithmic decision making (De Laat, 2018).

With the introduction of algorithmic decision making and the research that is so far done in it, a new debate started about whether algorithms are capable of making decisions without human interaction. Zarsky (2016) found that a popular critique against using algorithmic decision making is that it could be inefficient because it is an error-prone process. These errors can partly be explained by biases introduced by algorithms, such as social biases or technical biases embedded in the algorithm (Mittelstadt et al., 2016). A lot of people are however not aware that there are plenty of cognitive biases that influence human decision making (Bazerman & Moore, 2013). Humans also tend to make systematic and predictable mistakes and human decisions are also subject to bias (Zarsky, 2016). Humans tend to trust and accept the human decision making more than that of algorithms, without considering the amount of biases that influence the process.

Since there is no findable research done into the acceptance of algorithmic decision making, this research paper will try to map the field of social acceptance of machine learning in a decision making context. It was decided to only focus on machine learning, while algorithmic decision making on its own is a very broad field. This paper can be seen as an exploratory study mainly focused on creating a starting point for future research. The research question that will be used during this research will be:

"Is machine learning in a decision making context socially accepted?"

Decision making contexts can come in all forms, sizes, and degrees of complexity. It was therefore decided to investigate if social acceptance is influenced by context. For this research, different machine learning developments of the police are used as

an example of complex adaptive systems, since these developments are used with the retrospect of safety, but also privacy. Hypothesized is that social acceptance is influenced by the context. The following subquestion will be used during this research:

1. *Is social acceptance of machine learning situationally/context dependent?*

Since literature indicated that humans are not always aware of their own biases in a decision making context (Bazerman & Moore, 2013), it could be interesting to look at whether humans are aware of machine learning biases in a decision making context. The following subquestion is examined to find an answer to this question:

2. *How do perceived machine learning biases correspond to the biases found in literature?*

To further analyze how social acceptance is formed, it will be compared to the perceived biases of machine learning. The following question is used:

3. *Is there a correlation between social acceptance (q1) and the perceived biases (q2)?*

2. Related work

The theoretical framework will first go into biases introduced by humans. After that, it is shortly explained what machine learning exactly is to better understand where biases could come from. Then biases and dangers introduced by using machine learning in a decision making context are discussed. Finally, the current debate in whether algorithmic decision making should be used is mapped out.

2.1 Human biases in decision making

Bazerman and Moore (2013) introduce twelve different kinds of heuristics, or rules of thumb, which can be seen as bias by human decision making. These biases are sub-grouped into three categories. The first category of biases are biases emanating from the availability heuristic. This subgroup consists of two biases: the availability bias, which states that individual often judge events based on how they are readily available in memory, and the retrievability bias, which states that the judging of an event is biased by the memory structures of the individual.

The second subgroup of biases are biases emanating from the representativeness heuristic. This subgroup contains five different kinds of biases. The first one is insensitivity to base rates, the second one insensitivity to sample size and the third one misconceptions of chance. Besides those biases, this subgroup also contains the regression to the mean bias, which states that when speaking about subsequent trails, extreme events regress to the mean. Lastly, there is also the conjunction fallacy bias that states that people often falsely think that two events co-occurring are more likely than a greater set of events that are co-occurring. Tversky and Kahneman (1977) found similar biases and concluded that people tend to yield incorrect conclusions because they ignore the basic principles of probability.

The third and last subgroup of human biases are biases emanating from the confirmation heuristic, which consists of five different kinds of biases. The first bias is the confirmation trap, which states that people fail to find counterarguments and only search for confirmatory information for the arguments they believe in. Secondly, there is the anchoring bias, which states that people tend to make estimates based on an initial value, but tend to fail to adjust this estimate. The third bias is the conjunctive- and disjunctive bias,which states that people underestimate that disjunctive events can occur, and overestimate that conjunctive events can occur. The fourth bias is the overconfidence bias, which states that people are overconfident about the quality and rightness of their answers when answering difficult questions. The last bias is the hindsight and the curse of knowledge bias, which states that after an event occurred, people overestimate the degree in which they could have predicted the event beforehand.

Levinson (1995) states that, especially when performance is bad, people make recurrent errors. He found that when a complex system fails, people tend to take ill-adaptive measures to fix the system because they continue to seek confirmation of failing hypotheses. Next to that people tend to act immediately after a problem emerges, without doing an analysis first. Next to that Levinson also found biases similar to the conjunctive- and disjunctive bias and hindsight bias.

De Martino et al. (2006) looked into the effect of the framing effect on brain activity. The framing effect states that people's choices are remarkably sensitive to the way in which options are presented. The way a problem is presented can thus influence the way that people think about it. When a problem triggers parts of the brain that regulate emotional responses, the decision made is different than when not. They thereby include that emotions can play a big role when making decisions.

2.2 What is Machine learning?

To better understand the problem statement, a brief explanation of machine learning is provided. Machine learning is every method and technique that uses data to discover new patrons and that can generate models that can be used for effective predictions of the data. It can be seen as a research field that combines statistics, artificial intelligence, and computer science. Recommendation systems are one of the most popular applications of machine learning (Muller & Guido, 2016; Mittelstadt et al., 2016). Machine learning can be divided in regression (for example correlation) and classification (for example identifying spam) algorithms based on the prediction to be made. Learning can be accomplished by supervised learning (training and validating on labeled data) and unsupervised learning (looking for patterns) (Muller & Guido, 2016).

Using these methods, machine learning can be used for decision making by aggregating and processing data to for example optimize or classify certain outcomes. Difficulties in

understanding machine learning arise since most machine learning algorithms are black boxes. Input is given and a machine learning algorithm is chosen and tuned, which leads to an output. However the internal, underlying process which creates the algorithm is not opaque (Sametinger, 1997).

2.3 Machine learning biases in decision making

In this section, biases created by using machine learning are further elaborated. Friedman and Nissenbaum (1996) argue that there are three different biases can arise when using machine learning. The first way is when biases arise from the pre-existing social values found in the environment in which the algorithm is implemented. This can be intentional, for instance when the biases are embedded by the system designers, but it can also be an unintentional subtle reflection of values, for example when data is tagged by people. Next to that, it is possible that biases arise from technical constraints. A good example of a technical bias is when companies are listed alphabetically, and the companies at the top of the list get more business opportunities, just because they are on the top. Finally, it is possible that they arise from the emergent aspects of the used context. Emergent bias, for example, occurs when there is a shift of context and the end users of the algorithm change due to that the algorithm was originally designed for a different end user (and use purpose).

Mittelstadt et al. (2016) describe six concerns emerging from using algorithms when making decisions:

- Inconclusive evidence: It should be validated whether the algorithm actually finds a good statistical outcome, and does not find a false correlation between data.
- Inscrutable evidence: The connection between the data and the conclusion should be clear.
- Misguided evidence: It should be checked if the evidence is misguided. This is observer-dependent.
- Unfair outcomes: It should be monitored if the actions based on the algorithmic decision making are actually fair based on ethical criteria and principles. This is also observer-dependent.
- Transformative effects: Algorithms affect how people conceptualize the world. This can introduce ethical challenges.
- Traceability: Ethical assessment requires the responsibility and the cause form harm to be traced when a problem is found in one of the other five concerns.

These concerns can have serious consequences. Inconclusive evidence could, for example, lead to unjustified actions, especially when an algorithm is not validated and the cause of for instance a correlation is not known. Next to that inscrutable evidence could lead to opacity, since there is no transparent process that can explain why the algorithm makes certain choices because the algorithm is most of the time a black box. Misguided evidence can also result in one of the biases that Friedman & Nissenbaum (1996) discussed.

Besides consequences based on the evidence, there are also other consequences. Unfair outcomes can also lead to discrimination, especially when there is sensitive information in the data. Furthermore, transformative effects could lead to challenges for autonomy since “value-laden decisions made by algorithms can pose a threat to the autonomy of data subjects” (Mittelstadt et al., 2016, p. 9). Transformative effects can also lead to challenges for informational privacy since people do not want to be part of a group, and when they keep on being categorized in a group, they can ask for more informational privacy to prevent this. Finally, it is important that traceability should lead to moral responsibility. Computer programmers used to have the blame when an algorithm failed. Machine learning can, however, lead to black boxes. The programmers hereby don’t necessarily know what the algorithm does. It is therefore really important to use traceable methods to form the algorithm, so the harm can be traced back to the cause (Mittelstadt et al., 2016)

2.4 Human versus Machine Learning in a decision making context

It is currently still an ongoing debate about whether or not machine learning, and other algorithmic approaches, should be used in a decision making context. Zarsky (2016) found that a popular critique against using algorithmic decision making is that it could be inefficient because it is an error-prone process. These errors can partly be explained by biases discussed in the last section. As discussed in section 2.1 there are also a lot of cognitive biases that influence human decision making. Humans tend to make systematic and predictable mistakes and human decisions are also subject to bias (Zarsky, 2016).

Transparency is one of the arguments opposing algorithmic decision making since most algorithms are opaque due to the occurrence of black box mechanics (Mittelstadt et al., 2016; Zarsky, 2016). Creating transparency however is not preferred, since this could lead to a potential loss of competitive edge and privacy because if an algorithm is made publicly available they can be used by competition and can be wrongly interpreted (De Laat, 2018).

Next to the transparency debate, there is also an ongoing debate about whether the biases introduced by machine learning algorithms are more severe issues that the biases introduced by humans. It is argued that algorithms could make objective decisions, when programmed to do so or when the algorithms learned how to do so. Humans however always make subjective decisions, since bias is imprinted into our brains (Mittelstadt et al., 2016). McAfee et al. (2012) go further into how algorithmic decision making can be used in a decision making context in practice. They conclude that “data’s power does not erase the need for vision or human insight” (p. 7).

2.5 Summary

Since no previous (findable) research was done into the acceptance of machine learning or other algorithmic decision making tools, this could not be discussed in this section. There

are, however, as described in the previous sections, several dangers and biases introduced by both automated/human and non-automated/algorithmic decision making. A lot of people don't seem to be aware of human biases, but do use algorithmic biases as a counter-argument to not use machine learning in a decision making context. This research tries to find out whether people can point out machine learning biases and tries to find a relation between these biases and the social acceptance of these biases.

3. Method

This research examines the social acceptance of machine learning in a decision making context. The decision making context used here is a police setting. It was chosen to use a police setting due to its complex and adaptive characteristics. Machine learning application developments at the police are used with the retrospect of safety, but also privacy. A questionnaire was used to measure the social acceptance of machine learning (ML) in this setting.

3.1 Measurement

A questionnaire was used to collect the data. First, a brief introduction of the questionnaire is given, followed by an explanation of what machine learning is: an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed (Expert System, 2019). After that, the participants have to fill out some general information such as gender and age and are asked if they were already familiar with machine learning before reading the provided information.

The first part of the questions about machine learning are adapted questions of TAM3 (Venkatesh & Bala, 2008) and UTUAT2 (Venkatesh et al., 2003). These models both measure the acceptance of technology by using different criteria. Most of the questions can however only be asked when the respondents have used the technology. Since in this context it is not possible to let people use machine learning in a police setting, two criteria are extracted from TAM3 and UTUAT2: technology anxiety and performance expectancy. While technology anxiety can give insights in to why someone is hesitant about a technology, performance expectancy can give insights into why someone would use the technology.

To investigate if the acceptance of machine learning is affected by the scenario/decision making environment, different scenarios are outlined from actual implementations of machine learning in the policing sector. The scenarios are described in appendix 1. For each scenario, the opinion about trust, acceptance, and ethicality of machine learning is measured using statements such as "I trust machine learning in this setting". It was decided to measure the ethicality since Mittelstadt et al. (2016) found that people fear that machine learning algorithms do not act ethically. Trust was measured since Venkatesh and Bala (2008) found that trust can influence the acceptance of technology. This could both lead to an decrease in acceptance. A five-point Likert scale, strongly disagree (1) to strongly agree (5), with the anchor points as stated by Vagias (2006), can then be used to indicate in which extent the participant agrees with the statement. After that the participant is asked to indicate whether only a human, only machine learning or a mix of human and machine learning should be used to make decisions in the scenario.

After that, there follows an open question about the perceived dangers and biases in machine learning. This is asked to check whether the respondents are aware of the biases of machine learning. Participants could indicate whether they thought there were dangers when using machine learning and if yes, what the dangers are.

Lastly, there is a question about whether in general a respondent thinks that humans, machine learning or a mix of both should be used in a decision making context. For a full overview of the questionnaire, see Appendix 2.

3.2 Data collection procedure

Due to the exploratory nature of this study and the short time to conduct this study, it was decided to use students as a target population. To reach as many students as possible, an online survey was created in Google Forms. The survey consisted of the five sections described above: a general introduction into machine learning, general information, adapted TAM3 and UTAUT2 questions, the scenarios and the general question about biases and who should make the decision.

4. Results

4.1 Respondents

34 people filled out the survey. Two people did however not seriously fill out the survey, the results of 32 people were analyzed. 29 of the 32 respondents were male and 3 were female. The average age of the respondents is 21.4 (M=21.41, SD=3.046514, min=16, max=34). 26 of 32 were familiar with Machine learning, according to their opinion, before reading the explanation in the questionnaire.

4.2 Anxiety & performance expectancy

The mean and standard deviation of the statements about anxiety and performance expectancy are shown in table 1. Participants had to indicate on a Likert scale (strongly disagree (1) to strongly agree (5)) if they agreed with the statement. The first three statements cover the anxiety around the use of machine learning in a decision making context, the overall view of anxiety seems to be low, only the third statement (note that this statement is inverted) seems to indicate some anxiety. The other statements cover the usefulness. Overall the view on usefulness of machine learning in decision making seems to be positive.

Statement	Mean	Std. Deviation
Machine learning based decision making makes me nervous	1.81	0.821
Machine learning based decision making makes me feel uncomfortable	1.78	0.870
Machine learning based decision making does not scare me at all (inverted)	2.22	1.157
I think machine learning based decision can be useful to society	4.28	0.851
Using machine learning based decision making can increase the chances of achieving things that are important to society	4.22	0.832
Using machine learning based decision making can help society accomplish things more quickly	4.19	0.896
Using machine learning based decision making enhances the effectiveness of authorities	3.53	1.107

Table 1: The mean and standard deviation of the adapted UTUAT2 and TAM3 statements

4.3 Scenario analysis

The opinion about trust, acceptance, and ethicality of machine learning was measured on 5 scenarios with a 5 point Likert scale. A repeated measure ANOVA is used to evaluate the measurements. Acceptance($F(4) = 1.433$, $p = .003$) and ethicality($F(4) = 5.277$, $p = .003$) were rated differently between scenarios. This means that the values between groups are significantly different from within groups. To further analyze where the difference is occurring, a Bonferroni post hoc is used. This test found that the acceptance (mean diff = .719, $p = 0.033$) and ethicality (mean diff =1.000, $p = 0.001$) of scenario 1 "Classifying people" and scenario 4 "Cold cases" are significantly different from each other. The following estimates are used as evaluation:

- Acceptance; scenario 1 ($M = 3.625$, $SE = 0.205$) & scenario 4 ($M = 4.344$, $SE = 0.106$)
- Ethical; scenario 1 ($M = 3.188$, $SE = 0.208$) & scenario 4 ($M = 4.188$, $SE = 0.130$)

Trust is however not significantly different between scenarios with a mean of 3.531 in scenario 1 as the lowest value and 3.938 in scenario 4 as highest. There were no other significant differences found by comparing other scenarios. There were also no relations found between the ethicality and the acceptance (as described in Mittelstadt et al., (2016)) or between the trust and acceptance (as described in Venkatesh & Bala (2008)) within the scenarios.

When asked to choose between a human, machine learning or a combination of both as the best option in the scenario, the option of having both was the top answered question in all 5 scenarios,

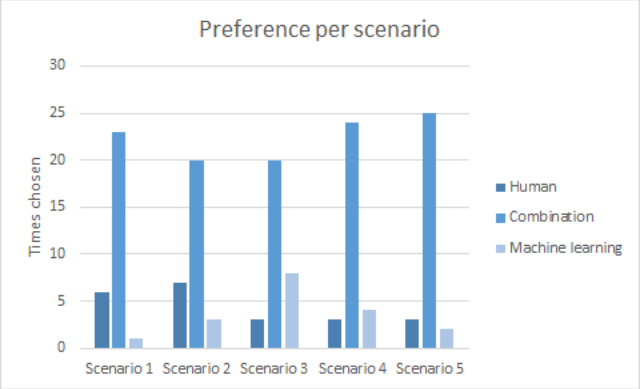

Chart 1: preferences per scenario

followed by a human (3/5 scenarios higher than only machine learning) (Chart 1).

4.4 Correlation social acceptance and perceived biases

To answer subquestion three (Is there a correlation between social acceptance and the perceived biases?) it was tried to find a relation between social acceptance and the perceived bias. The perceived bias was therefore transformed into a binary variable, with value 0 if the respondent did not think there were dangers or biases and value 1 when a respondent thought there were biases or dangers.

An independent samples t-test was used to search for a relation. The anxiety, perceived usefulness as well as the individual scenario acceptance were used. There is no significant relationship found between any of these variables and if there are biases perceived.

4.5 Open questions

The open question can give insight into whether or not the respondents are aware of biases and dangers of machine learning without educating them about potential biases and dangers. The answers to the open questions are manually labeled and categorized according to the biases (social, technical and emergent) as stated by Friedman and Nissenbaum (1996).

11 of 32 people (34.4%) are not aware of the dangers of biases that decision making based machine learning introduced. From the people that did mention biases, 19 mentioned technical biases and 2 mentioned social biases.

The technical biases mentioned were mostly focused on whether the algorithm is able to predict with a 100% accuracy, which was mentioned 6 times. One of the respondents mentioned that he thought that wrong decisions were a danger. He mentioned that: "Machines are never 100% correct, this can also be seen at AI [applications]. If heavy penalties or human lives are involved, you must be 100% sure that it is the correct choice. I think that, at the moment, there is not always such a certainty". Besides that it was mentioned that "It is a system and people cannot always be

categorized, that's why machine learning will never work 100% and (unnecessary) mistakes can be made." Next to that, it was mentioned three times that machine learning outcomes should be checked and that the human should make the concluding decision. Another respondent agreed on this and mentioned that it could be hard to detect exceptional cases, which could, for example, mean that criminals are not caught. It was also mentioned that people could find loopholes to avoid the machine learning. Other technical biases that were mentioned are that verification of an algorithm can be hard and that the algorithm is dependent on correct data. Finally, it was mentioned that it is not possible yet to combine hardware and software into a combined consciousness and that "The lack of such consciousness, and thereby also the lack of empathy and perspective, can lead to "cold" (but often completely logical) decision-making. A child who steals an apple can be automatically labeled as a suspect by machine learning. After all, it is criminal behavior. However, a human would only make this decision after he/she had wondered whether the child might not get food at home."

Two respondents mentioned a social bias. The first one mentioned that "The system is dependent on data that is added. This can be misused". This can be seen as a social bias since the misuse/tampering with the data could be seen as a social influence on the algorithm. Another social bias that was mentioned is: "I think there is a danger that machine learning will be used for something good such as safety, but certain liberties will be sacrificed for that. Take China, for example. That country is full of cameras where people keep an eye on society. And this country is trying hard to steer society. Machine learning makes it easier for the population to monitor. After which they can steer the population even more effectively". The preexisting social values of the Chinese government are here used in the machine learning algorithm, which has great consequences.

A full overview of the obtained answers, labels and categories is documented in Appendix 3 & Appendix 4.

5. Conclusion

From the respondents, the majority is positive about the general usefulness of machine learning and is not afraid of it. Trust in machine learning between the different scenarios is semi-positive to positive. Remarkably, when comparing ethicality and acceptance, they are affected by the scenarios. Scenario 1 "Classifying people" has a significantly lower acceptance and is perceived as less ethical than using machine learning in the selection of cold cases (scenario 4). This might suggest that acceptance is decreasing when machine learning is used on more personal and ethically loaded topics. There is however further research needed to confirm or deny this. The first subquestion*: Is social acceptance of machine learning situationally/context dependent?* is thereby answered, it seems like there could be a difference in acceptance between different scenarios.

However, when confronted with which of machine learning, humans or a combination can be best used to make decisions, the combination of machine learning and humans is highly favored. This can suggest that machine learning in combination with a human (for example as a supervisor) is trusted and the advantages of using the best of both advantages outweigh the disadvantages. Again, there is further research needed to confirm or deny this suggestion.

The human biases and machine learning biases are introduced in the theoretical framework. Remarkably most of the respondents actually mentioned a danger or biases that could be introduced when using machine learning in a decision making setting. It was possible to classify most of these mentioned biases and dangers into two of the three categories of biases that Friedman and Nissenbaum (1996) mentioned: social and technical biases. Most of the respondents could name a technical bias. Interestingly enough most of the named biases and dangers correspond with dangers described in Mittelstadt et al. (2016). Emergent biased were not mentioned. No one mentioned that human biases also exist. The third subquestion*: How do perceived machine learning biases correspond to the biases found in literature?* is thereby answered, people tend to detect technical biases, have harder times detecting social biases and do not detect human or emergent biases.

To answer the third question, *Is there a correlation between social acceptance (q1) and the perceived biases (q2)?*, it was tried to find a relation between the perceived biases and social acceptance. There were no relations found.

Conclusively, machine learning seems to be accepted and trusted in most of the scenarios, which supports the trust in and acceptance of machine-learned decision making. However, acceptance and trust might decrease as the employment of machine learning in decision making touches more personal or more ethical subjects. People seem to have a general understanding of the biases of machine learning in a decision making context. There seems to be no relation between these biases and social acceptance.

6. Discussion

Convenience sampling was used in the quantitative part of this study. This means that the respondents were easily accessible, they were mainly found by spreading the survey using Facebook and WhatsApp. The respondents are most likely not a good reflection of society since it was not a random sample (Burns & Burns, 2008). This was visible by the fact that the average age is not the average age of the total Dutch population and the number of males and females was also not representative. This decreases the ability to generalize the results to the whole population.

Additionally, the labeling of the open questions, to match the biases of Friedman and Nissenbaum (1996), is done manually and therefore open for interpretation. It might occur that a category is wrongly labeled in someone else's opinion. However, this has no further impact on the research.

Finally, this research consisted of a small quantitative and a small qualitative study, due to the explanatory nature of this study. This decision, however, brought some limitations with it, since it was not possible to find causes for differences or to ask further into the biases. Since there was no findable research into the acceptance of using machine learning in a specific context, the validity of the questions of the survey is low, since existing methods could not be used. It was therefore tried to create the survey partly based on TAM3 and UTAUT2 questions, and based on statements from literature. It is however not possible to validate the results. To create a more expressive research, more elaborated research into social acceptance is needed.

7. Future work

This paper was mainly an exploratory study into the social acceptance of machine learning in a decision making context. Based on this study, some interesting starting points for future research are mentioned below:

- This research focused on scenarios in a police setting. There are however a lot of other interesting research fields where social acceptance can be different than in this setting. For example in the healthcare field, where machine learning is also used a lot.
- This research consisted of a small quantitative and qualitative part. It could, however, be interesting to do a larger qualitative study into how much people know about the biases. Even though most people in our survey could name a danger of bias of a machine learning algorithm, it could be a good idea to do further research into which biases people heard of and if they are also aware of the human biases in decision making.
- A significant difference in acceptance was found between the first scenario, the classification of people, and the fourth scenario, the use of machine learning to decide which cold cases to solve. Due to the quantitative manner of this research, it was not possible to find a cause for this difference. It could be interesting to do further research into why people think different about these scenarios. It could, for example, be the case that acceptance is decreasing when machine learning is used on more personal and ethically loaded topics.
- It was found that the combination of a human and a machine learning algorithm seems like the optimal decision making situation. Due to the quantitative nature of the survey, it was not found why this is the case. The perceived biases indicated that machine learning in combination with a human (for example as a supervisor) is trusted and the advantages of using the best of both advantages outweigh the disadvantages. Since it was not possible to confirm or deny this in this research, it could be an interesting topic to do further research into.
- Mittelstadt et al., (2016) described a relation found between the ethicality and the acceptance. Venkatesh and Bala (2008) described a relation between trust and acceptance. These relations were not found in this research. It could be interesting to look further into these relations, and check whether these relations are present in other decision making contexts.
- Since there exists a black box, there is uncertainty that the algorithm is ‘smart’ enough to process cases correctly that are unique or have not occurred before. As described in Mittelstadt et al., (2016) there is therefore always a certain risk in misinterpretation or erroneous/incomplete data. A classic example for this is that a machine learning algorithm predicting vandalism change for an individual might expect that blue-eyed people are more likely to be vandals, simply because in the data vandals are predominantly blue-eyed people. Thus, it is hard, if not impossible, for a machine to clarify the causality of cases. This means that a machine might never be fully trusted in decision making where a decision could have a “serious impact on someone's life” as stated by a respondent. It would be interesting to further investigate where the line is drawn between trusting an algorithm and questioning the black box.

REFERENCES.

- Bazerman, M., & Moore, D. A. (2013). Judgment in managerial decision making.
- Burns, R. P., & Burns, R. (2008). Business research methods and statistics using SPSS. Sage.
- Danaher, J., Hogan, M. J., Noone, C., Kennedy, R., Behan, A., De Paor, A., ... & Murphy, M. H. (2017). Algorithmic governance: Developing a research agenda through the power of collective intelligence. Big Data & Society, 4(2), 2053951717726554.
- De Laat, P. B. (2018). Algorithmic decision-making based on machine learning from Big Data: Can transparency restore accountability?. Philosophy & Technology, 31(4), 525-541.
- De Martino, B., Kumaran, D., Seymour, B., & Dolan, R. J. (2006). Frames, biases, and rational decision-making in the human brain. Science, 313(5787), 684-687.
- Domingos, P. (2015). The master algorithm: How the quest for the ultimate learning machine will remake our world. Basic Books.
- Expert system (2019). What is Machine Learning? A definition - Expert System. Retrieved March 12, 2019, from https://www.expertsystem.com/machine-learning-definition/
- Flach, P. (2012). Machine learning: the art and science of algorithms that make sense of data. Cambridge University Press.
- Friedman, B., & Nissenbaum, H. (1996). Bias in computer systems. ACM Transactions on Information Systems (TOIS), 14(3), 330-347.
- Kahneman, D., & Tversky, A. (1977). Intuitive prediction: Biases and corrective procedures. Decisions and Designs Inc Mclean Va.
- Levinson, S. C. (1995). Interactional biases in human thinking. In Social intelligence and interaction (pp. 221-260). Cambridge University Press.
- Marr, B. (2018). How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read. Retrieved March 10, 2019, from https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/
- McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D. J., & Barton, D. (2012). Big data: the management revolution. Harvard business review, 90(10), 60-68.
- Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. Big Data & Society, 3(2), 2053951716679679.
- Müller, A. C., & Guido, S. (2016). Introduction to machine learning with Python: a guide for data scientists. " O'Reilly Media, Inc.".
- Sametinger, J. (1997). Software engineering with reusable components. Springer Science & Business Media.

- Vagias, W. M. (2006). Likert-type scale response anchors. clemson international institute for tourism. & Research Development, Department of Parks, Recreation and Tourism Management, Clemson University.
- Venkatesh, V., & Bala, H. (2008). Technology acceptance model 3 and a research agenda on interventions. Decision sciences, 39(2), 273-315.
- Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. MIS quarterly, 425-478.
- Zarsky, T. (2016). The trouble with algorithmic decisions: An analytic road map to examine efficiency and fairness in automated and opaque decision making. Science, Technology, & Human Values, 41(1), 118-132.

Appendix 1 overview scenarios

Scenario 1: *Classifying people*

Machine learning can be used to classify people into groups (Dees, 2018). This means that people with for example the same ethical background and level of education will be in the same group. After the classification, the groups can have an assigned risk level. Based on this risk level the police can put extra surveillance on people.

One real-life example of the classification is the Top600 that Amsterdam created (Gemeente Amsterdam, 2019). This list contains people who have a high impact on delicts in the past. Amsterdam afterward informed all 600 people on the list why they were on there and assign everyone a personal advisor to help find a fitting job or education.

Scenario 2: *Surveillance analysis*

There are a lot of cameras in public areas used to monitor activity. Currently, most of those cameras are monitored by people. There are however machine learning technologies coming up that make it possible to automatically detect actions in videos (Dees, 2018; Lalawat, 2018). This can make it possible to spot crimes or every other desirable action, automatically.

One real-life example of this could be that selling drugs can be spotted on festivals when the terrain installed cameras. An example is the Dutch "Freshtival" festival (Middelbos, 2018).

Scenario 3: *predicting home burglary*
One of the uses of machine learning is predicting whether certain homes are more or less prone to burglary, using crime rates and historical data in a neighborhood.

A real-life example is the so-called 'Criminality Anticipation System from the Dutch police force ("Data Science predictive policing", 2019). This system uses available data to color districts on a map, creating a map with the possibility of a crime in a certain region. This is analyzed by task forces and used to distribute available officers.

Scenario 4: *Cold cases*

A cold case is an unsolved police case with a minimum sentence of 12 years and no progress in the last three years. The Dutch police is working on an artificial intelligence implementation with machine learning to reselect these cases (van der Beek, 2019; ("Nederlandse politie gebruikt AI om moeilijke zaken op te lossen - Techzine.nl", 2018). New data is used to evaluate the cold cases and cases that are more likely to be solved with the new data is selected. Evaluating a cold case manually could take weeks, while this implementation can potentially evaluate the cold case in about a day.

Scenario 5: *Sorting businesses based on data*

A ranking system can be created to assess more fraudulent companies. This can be based on data of income and expenses, or on data such as location and the kind of business. The police can utilize these rankings to determine which businesses are more necessary to be investigated.

References:

- Data Science predictive policing. (2019). Retrieved from https://it.kombijdepolitie.nl/predictive-policing
- Dees, T. (2018). How deep learning is transforming police investigations. Retrieved from https://www.policeone.com/police-products/police-technology/software/video-analysis/articles/476485006-How-deep-learning-is-transforming-police-investigations/
- Gemeente Amsterdam. (2019). Top600. Retrieved from https://www.amsterdam.nl/wonen-leefomgeving/veiligheid/top600/
- Lalawat, P. (2018). Deep Learning Is a Blessing to Police for Crime Investigations | Analytics Insight. Retrieved from https://www.analyticsinsight.net/deep-learning-is-a-blessing-to-police-for-investigations/
- Middelbos, J. (2018). Twaalf megacamera's speurden dit Freshtival Het Rutbeek af - Via Certus. Retrieved from https://www.viacertus.nl/cameratoezicht-freshtival/
- Nederlandse politie gebruikt AI om moeilijke zaken op te lossen - Techzine.nl. (2018). Retrieved from https://www.techzine.nl/nieuws/404929/nederlandse-politie-gebruikt-ai-om-moeilijke-zaken-op-te-lossen.html?redirect=1
- van der Beek, P. (2019). Politie positief over inzet ai bij cold cases. Retrieved from https://www.computable.nl/artikel/nieuws/big-data/6599469/250449/politie-positief-over-inzet-ai-bij-cold-cases.html

Appendix 2 The questionnaire

Screen 1:

This questionnaire is part of a study for the course Decision Making in a Complex System Setting at the UvA. The purpose of the research is to map out how people view technological developments in the Machine Learning area. The answers to this survey are completely anonymous and are not shared with third parties.

Screen 2:

- Age
- Gender
    - Man
    - Woman

Screen 3:

Machine learning consists of three steps:

- A user has data with which he / she wants to predict something or that he / she wants to categorize. The user gives this data as input to a computer.

- The computer creates a model. The user can adjust this model slightly by adjusting parameters (which means that the user can indicate which elements are important and which are not).

- The computer gives a prediction or category label as output

Example: Weather data

Sensors collect data every day. A weather forecaster gives this data to the computer. The computer generates a model based on historical data and searches for patterns in it. The computer thus learns, for example, that a high pressure area provides heat in the summer. Thus, when the sensors measure a high pressure range, the computer outputs high temperatures.

Were you familiar with machine learning before this introduction?

- Yes
- No

Screen 3:

Indicate how you feel about the following statements based on the 5 point likert scale:

- Machine learning based decision making does not scare me at all
- Working with machine learning based decision making makes me nervous
- Machine learning based decision making makes me feel uncomfortable
- I think machine learning based decision making can be useful for society
- Using machine learning based decision making can increase the chances of achieving things that are important to society
- Using machine learning based decision making can help society accomplish things more quickly
- Using machine learning based decision making enhances the effectiveness of authorities

Screen 4:

The scenarios from appendix 1 are described here. All the scenarios are followed up by the following statements and question, which again could be answered using the Likert scale:

- I trust the use of machine learning in this scenario.
- I accept the use of machine learning in this scenario.
- I think using machine learning is ethical in this scenario.
- I think that this decision can be best made by:
- A human

- Machine learning
- A mix of both

Screen 5:

Do you think there are dangers when Machine learning is making these kinds of decisions? If yes, what do you think that these dangers are?

In general I think that machine learning can:

- can make decisions on it's own
- should only be used as a support tool, humans should still be involved in the decision making
- should not be involved in decision making processes

Appendix 3 labeled answers open question “Dangers”

Answer	**Label**
Foutieve keuzes. Machines zijn nooit 100% correct, dit is bij AI ook te zien. Als er zware straffen of mensenlevens in het spel zijn moet je 100% zeker weten dat het een correcte keuze is. Ik denk dat er, op dit moment, niet altijd een dergelijke zekerheid is.	not 100% correct
Nee	no
Nee	no
Mensen gaan trucjes proberen te verzinnen om de machine learning te ontlopen	people find loophole
Nee	no
Ja, het kan bijvoorbeeld uitzonderingsgevallen niet detecteren, denk ik. Waardoor misdadigers toch niet betrapt worden bijvoorbeeld	not detecting exceptions
Fouten zijn makkelijk te maken omdat het systeem niet optimaal kan aanpassen	dependant on correct data
Een beslissing is nooit 100% zeker, het is daarom handig om menselijke controle hierover te hebben	not 100% correct
Nee	no
Nee	no
Nee	no
Nee	no
ML kan ook fouten maken en als de uitkomst standaard als feit wordt gezien en data nog niet eerst gecontroleerd wordt kan dat grote problemen opleveren.	not 100% correct
False positives	False positives
Het kan bij scenario 2 bijvoorbeeld leiden tot een verkeerde focus	Wrong focus
onterechte beschuldigingen	False positives
Voornamelijk in het geval waarin de keuze wordt gemaakt door enkel machine learning, dit kan beter als data worden behandeld die wordt gecontroleerd voordat deze wordt gebruikt	dependant on correct data
Gevaar voor blind vertrouwen zonder menselijke verificatie	Verification is needed
Een mens kan op bepaalde oogpunten toch elders tegen aan kijken	Point of view

Zodra AI bezwaard wordt met vraagstukken en/of beslissingen welke direct negatief invloed kúnnen hebben op iemands (kwaliteit van ...) leven, ontstaat gevaar zodra aan deze besluitvorming niet getwijfeld wordt. Hoewel we inmiddels vér voorbij de Turing-test zijn, is het nog niet gelukt om software en hardware dusdanig te combineren dat er een bewustzijn ontstaat. Sterker nog, we weten niet eens wát het precies is. Het ontbreken van een dergelijk bewustzijn én daarbij dus ook het ontbreken van inlevings- en relativeringsvermogen kan zorgen voor "kille" (doch vaak volstrekt logische) besluitvorming. Een kind dat een appel steelt, kan door machine learning automatisch worden gelabeld als verdachte. Het is immers crimineel gedrag. Een mens zou echter dit besluit pas nemen nádat hij/zij zich heeft afgevraagd of het kind misschien thuis geen eten krijgt.	Lacks insight
Ik denk dat het gevaar bestaat dat machine learning wordt gebruikt voor iets goed zoals veiligheid maar daarvoor worden bepaalde vrijheden opgeofferd. Neem bijvoorbeeld China. Dat land staat vol met camera's waar men de samenleving in de gaten houdt. En dit land probeert al de samenleving hardhandig te sturen. Door machine learning kan de bevolking nog makkelijker in de gaten houden. Waarna ze nog effectiever de bevolking kunnen sturen.	Machine in control
Dat er onjuiste beslissingen worden gemaakt door bv. foutieve input data of onregelmatigheden, de mens zal toch samen met de machine learning uitkomst een eindbeslissing moeten maken	not 100% correct
Nee	no
Nee	no
Het is een systeem en mensen zijn niet altijd in hokjes of een systeem te duwen, waardoor het machine learning nooit 100% zal werken en er (onnodige) fouten gemaakt (kunnen) worden.	not 100% correct
Er kunnen fouten gemaakt worden of dingen vervalst	not 100% correct
Het systeem is afhankelijk van toegevoerde data als er iets wordt gedaan wat nog niet bekend is in het systeem kan hier makkelijk misbruik gemaakt van worden.	misuse
Veel dingen moeten door mensen worden bepaald	Verification is needed
Als machine learning veel fouten zou maken levert dit uiteindelijk juist meer werk op omdat er veel dingen misschien onderzocht worden terwijl dit niet nodig zou zijn, daarom denk ik dat er vooral in het begin altijd ook een persoon mee moet kijken of iets dergelijks	Verification is needed
Nee	no
Een computer alleen een beslissing laten maken zonder dat een mens hier inzicht in heeft blijft in mijn ogen altijd tricky	Verification is needed
Nee	no

Appendix 4 Open question categories

COUNT	UNIQUE TAGS	CATEGORY
11	**no**	**not labeled**
4	**Verification is needed**	**technical**
6	**not 100% correct**	**technical**
1	**people find loophole**	**technical**
1	**not detecting exceptions**	**technical**
2	**dependant on correct data**	**technical**
2	**False positives**	**technical**
1	**Wrong focus**	**technical**
1	**Machine in control**	**social**
1	**Point of view**	**technical**
1	**Lacks insight**	**technical**
1	**Misuse**	**social**