Algorithmic appreciation or aversion: does the representation of an algorithm change the trust placed in it?

Sander de Jong (s2096943)
Graduation Thesis
Media Technology MSc program, Leiden University
Thesis Advisors: Maarten H. Lamers & Maarten W. Bos
sander-dejong@hotmail.com

August 28, 2020

Abstract

Research on algorithms shows that people tend to overtrust computers but react strongly to mistakes, dismissing the system completely, and subsequently trust their own judgement more. People lose far less trust when they see humans make similar mistakes. We test this phenomenon - known as algorithmic aversion - in this study. Participants perform several estimation tasks. They are provided with an algorithm's estimated answers as suggestions. During the experiments, participants see that the algorithm is not flawless and errs substantially. We use the reactions to these errors in an attempt to reproduce the algorithmic aversion effect. Two additional variables are taken into account: the representation of the algorithm and the participants' affinity for technology. The representation is tested by providing one group with information about the workings of the algorithm and its accuracy, while the other performs the tasks without receiving this information. The results show that the algorithmic aversion effect occurs in the representation group in both experiments. This implies that people place trust in the proposed accuracy and form of an algorithm, but quickly lose this trust when an estimation is outside the promised accuracy boundaries. We find a similar effect in the group of participants who perform better than the algorithm. They deviate significantly more from it after they see the difference in performance between them and the algorithm, leading to lower scores as the algorithms are quite accurate on average. Participants deviate significantly more from the algorithms for suggestions that are outside of the promised accuracy range. This implies that they use the algorithms as an anchoring tool, but are still critical of their suggestions. Future work should elaborate on this effect as it is a hopeful sign for hybrid intelligence, where the algorithm augments the human decision process instead of replace it. The level of technical affinity does not have a significant influence on trust.

1 Introduction

Imagine a society in which government employees are replaced by artificial intelligence (AI). They determine whether farmers get government subsidies, which schools you are allowed to send your child to, when to see the doctor, and they even adjudicate simple legal disputes. This might seem a futuristic scenario, but Estonia has already deployed AI in 13 different government tasks and is currently working on an 'AI judge'[1]. While Estonia is a pioneer in this regard, other countries use AI systems to assist with important decisions in, for example, health care[2][3], in self driving cars[4] and in the courtroom[5][6]. It is only a matter of time before similar systems like the ones in Estonia are deployed in other countries. AI systems do and will determine the outcome of important decisions in our lives.

Work that was done by humans for years has now been taken over by computers. Apart from the obvious impact on the job market, it is interesting to see what the reaction of the citizens will be to such radical changes. While a human expert can often be asked what the reasoning is behind the decision, many AI systems are black box systems, in which the precise algorithmic weights are hidden from view

Currently most AI systems are used as assistants to the expert that is responsible for the final decision. In the future, more countries might follow in the footsteps of Estonia and make these systems autonomous. To what extent the outcomes of the algorithm will still be checked by experts depends on the implementation in these countries. Advisory or autonomous, in either case these systems will have a major impact on important decisions regarding people's lives. Research into the acceptance of these systems is lacking. A lot of related topics have been studied which provide a good theoretical basis to start with, but more research is needed that tests what factors influence the trust in algorithms. This study researches the effect of two factors on the people's trust: the representation of the algorithm and the technical affinity of participants.

The algorithms used in the experiments in this study advise people instead of deciding autonomously. They are exemplary for most algorithms in use today and also allow for evaluating what people decide to do with the information instead of merely accepting or rejecting the autonomous algorithm's suggestion.

Research on the stance of humans towards algorithms shows that people tend to overtrust computers but irrationally react strongly to mistakes, dismissing the algorithm completely, and subsequently trust their own judgement more[7]. This is an irrational reaction, because it also happens on tasks which computers have been proven to be better at on average. When people see humans make similar mistakes, they are a lot more accepting of them, losing far less trust in their judgement than they would if it was a computer. Dietvorst coined 'algorithmic aversion' as term to describe this effect in 2015, which is often used in related research.

This study tests whether the way an AI system is presented has an influence on the algorithmic aversion effect. Participants are asked to perform estimation tasks that have also been completed by an algorithm. The algorithm's answers are presented as suggestions to the participants. In addition to these suggestions, one group gets to see a representation of the algorithm (from now on called the representation group). The goal of the research is to examine whether getting more information about an algorithm influences the trust placed in the system.

It also tests the algorithmic aversion effect for people with varying technical affinity. If it is true that seeing an algorithm make a mistake makes one distrust the system, people who have more experience with computers have seen algorithms err more and might be more sceptical about them than people who have less experience. Furthermore, the algorithmic aversion effect might be less pronounced for the former group during the experiment because they already had less trust in algorithms to begin with or have learned to be more tolerant of mistakes for tasks where on average algorithms perform well.

Knowing what influence the presentation of an algorithm has on the trust placed in it is crucial knowledge in the adoption of AI systems. People might be more willing to use the systems when they are given the right information. If the algorithmic aversion effect is indeed more profound for people with low technical affinity, developers might be inclined to manage people's expectations of the algorithm better to prevent this shift from happening.

The paper is structured in the following way: Section 2 discusses related work including the psychological background of decision-making processes, the relation between humans and algorithms, different measures of the technical affinity of people and the likelihood that people conform to decisions made by computers. Section 3 explains the methods used and the difficulties faced in conducting this experiment during the COVID-19 pandemic. We describe the statistical analyses to test the hypotheses for the two experiments in their respective results sections. The discussion section elaborates on these answers and describes future work.

2 Related work

This section gives an overview of factors involved in human decision processes and misconceptions about them. We discuss the relationship between humans and algorithms to give a theoretical background for the hypotheses which are tested in this study. We outline the scales used to measure technical affinity and discuss theories on conformity to computers.

2.1 Human decision processes

Humans tend to have an explanation for their decision processes, often attributing it to their consciousness. Dijksterhuis argues that the subconsciousness is often undervalued and is in some cases more important than consciousness in complex decisions such as the ones mentioned in the introduction[8]. It is important to make the distinction between simple and complex decisions. With simple decisions the factors to be considered can still be processed by the brain, and the precision it gives leads to critical and sound decisions. Complex decisions involve too many factors to consider all the ramifications of the decision. The subconscious can handle a lot more information and is therefore able to process an abundance of options.

To test the importance of subconsciousness in complex decisions, Dijksterhuis let participants choose between different alternatives. The participants were divided in three groups. The first had to choose immediately, the second group got to take a short look at the problem but then was distracted by a task for a few minutes before making a decision and the third group got a few minutes to analyse the possibilities before deciding. The first group performed the worst as expected, but the second group outperformed the third one. The method of the second group is equivalent to "sleeping on" an important life decision, which is often considered irrational, but has now shown to work for some complex problems. Dijksterhuis and colleagues call this the 'deliberation without attention effect'[9], which describes that the subconsciousness is sometimes able to make better decisions because it can process the large amount of information better than our conscious mind.

This deliberation without attention effect cannot be conveniently placed in a dual process model. A dual process model proposes two systems of thinking. Kahneman calls them System I, or the 'intuition' module, and System II, or the 'reasoning' module[10]. System I is described to be fast, intuitive, emotional and subconscious, while System II is slow, logical, calculating and conscious. Slow unconscious decision processes such as the ones described above do not fit in either of the systems.

Humans also tend to think their decisions are consistent while it has been shown that they often are not. Studies in the United States and the United Kingdom showed that judges do not give the same judgement for identical cases twice, sometimes not even beyond chance level[11][12]. It is clear that human decision processes are not perfect and are prone to errors.

These examples are useful for our research because people have to choose between their own judgement and an algorithm's while they often misinterpret the way human decision processes work. Moreover, most participants probably have limited knowledge of algorithms (assuming most of them are not computer scientists). This combination of factors could be an explanation for the irrational mistrust of algorithms after they err described in the algorithmic aversion effect.

2.2 Anchoring effect

Human estimations based on suggestions, as the ones in this study, are biased by the anchoring effect[13][14]. Whenever people are presented with suggestions for a decision, they take this as anchor point and have a hard time evaluating the situation without this added information. This bias leads to irrational decisions because people stay too close to this anchored value when making an estimation.

While estimation tasks such as estimating the number of jelly beans contained in a jar are very old and research on them dates back to the year 1906[15], these are often studied with the wisdom of crowds in mind. In studies based on the wisdom of crowds principle, a large number of estimations are used to get to the final answer. This large amount of estimations cancels out some of the individual biases that might occur. In the task in this research, participants are faced with just one suggestion, making the anchoring effect far more impactful.

2.3 Algorithmic appreciation leading to algorithmic aversion

When given a choice between a human and computer as manager, participants chose the computer for analytical tasks, deeming them better at it[16]. For social tasks they preferred humans, considering computers less fair, less trustworthy, and more likely to evoke negative emotion for tasks that people think require uniquely human skills. Herz also showed that people trust the analytical skills of computers[17]. In his experiment, participants had to choose a human or computer advisor to either perform an analytical task or a social one. When the task at hand was unknown, participants chose the human advisor. If it was known beforehand that it was an analytical task, they chose a computer. This shows that people trust computers only on specific tasks and are particularly sceptical about their social capabilities. Castelo and colleagues found that perceived objectiveness of the task positively influences the willingness of people to offload it to an algorithm[18]. This can be mitigated by increasing the affective human-like qualities of the algorithm. Logg and colleagues showed that people choose the forecast of an algorithm instead of their own, unless they are an expert in the particular field[19]. People are less likely to reject their own forecasts than those of others, but in both cases prefer the algorithm.

Dietvorst and colleagues tested whether people trust humans or computers more when it comes to forecasting[7]. The tasks that were performed by both algorithms and people in these experiments also fall under the analytical tasks described in the previous paragraph. For different forecasting tasks participants had to choose between a human forecast (either their own or another participant's, depending on the experiment group) and the forecast made by the computer model. While at first participants relied on the model, this changed after they were presented with the errors of the model. This loss of trust was not noticeable when presented with similar human errors. He calls this algorithmic aversion: the reluctance to use a computer model when people have seen it err.

In an attempt to find a possible solution to this aversion, Dietvorst and colleagues showed that people are more likely to use an algorithm when they are allowed to change its forecasts[20]. Even after errors of the algorithm were shown, people kept using the algorithm and had more confidence in the algorithm than in their own performance while the control group (who could exclusively use their own estimations or the algorithm's suggestions) had more confidence in their own performance. Interestingly, the confidence in the algorithm's performance was similar when people were able to adjust it by only 10 percentile instead of freely. This suggests that people trust computers more when they have some control over the outcome. Restricting participants to changing it by only 10 percent also improved their performance since the algorithm is often better at predicting and changes generally only make it worse.

Kizilcec showed that people trust an algorithm less when their expectations of it are not met, unless they are given an insight into the algorithm through explanation[21]. However, adding outcome-specific information to the explanation to further increase transparency undermined the positive impact.

2.4 Technical affinity

This study tests the algorithmic aversion effect for people with varying technical affinity. Logg and colleagues showed that the expertise of people on the subject matter of a task influences the trust they have that an algorithm can outperform them in that particular task [19]. We are curious whether the technical affinity of the participants has an influence on the trust in algorithms. To test this, an appropriate measurement for technical affinity is needed.

While a definitive scale for technical affinity is missing, it has been researched from multiple perspectives. Edison and Geisler found several attributes contributing to technical affinity: dispositional optimism, need for cognition and self-efficacy[22]. Need for Cognition (NFC) is a term introduced in the 1950s to describe the tendency for an individual to engage in and enjoy thinking[23]. This attribute is relevant for this study because Cacioppo and Petty found that people with a low NFC tend to have difficulties solving complex problems[23]. It is likely that they want to offload these tasks to an algorithm.

In recent years several new methods to measure technical affinity have been developed, ranging from evaluating 'geekism' ("the need to explore, to understand and to tinker with computing devices"), computer anxiety and all-round tests for technical affinity[24][25][26]. All these are self-assessment questionnaires and most of them are quite long. In this study, the Affinity for Technology Interaction (ATI) scale is used, introduced by Franke and colleagues[27] as a more concise test, using the psychological construct of NFC as base. It is a self-assessment questionnaire consisting of nine questions which use a 6-point Likert scale. These result in a mean score that can be used as measurement of technical affinity. For a list of all the questions in the ATI see Appendix A.

As all self-assessment surveys, the ATI is in danger of falling victim to the Dunning-Kruger effect[28]. People who are not an expert in a certain field but have some knowledge about it often overestimate their knowledge. This might hinder the variance found while using this scale and is something to take into account while evaluating the results of the test.

2.5 Conformity

Due to the emergence of AI assistants in phones, cars, and electronic devices at home, much research has been conducted on the interaction between humans and these assistants. This followed up on earlier work done on human-robot interaction. Some of these studies focused on the question whether humans conform to robots the same way they do to groups of people.

Eventually it was thought that this was not the case, but Salomons concluded that people do conform to robots if the answer to the question is ambiguous instead of clear[29]. When an answer is given that is clearly incorrect, people don not conform to robots while they do in the same circumstances to humans. The fact that people do conform when answering ambiguous questions has to do with trust, according to the authors. When the trust in the capabilities of the robot is higher than the trust in their own judgement, people do conform to robots. However, when this conformity leads to an incorrect answer, people will not conform again. This is in line with the conclusions of section 2.3.

Conformity is interesting in the context of this research because it tests how people value the decisions made by computers. However, conformity was tested with embodied machines in different forms such as robots or voice assistants. Humans might be more likely to conform to these, although it has been shown that the human-like qualities of the robots do not influence the rate of conformity[30].

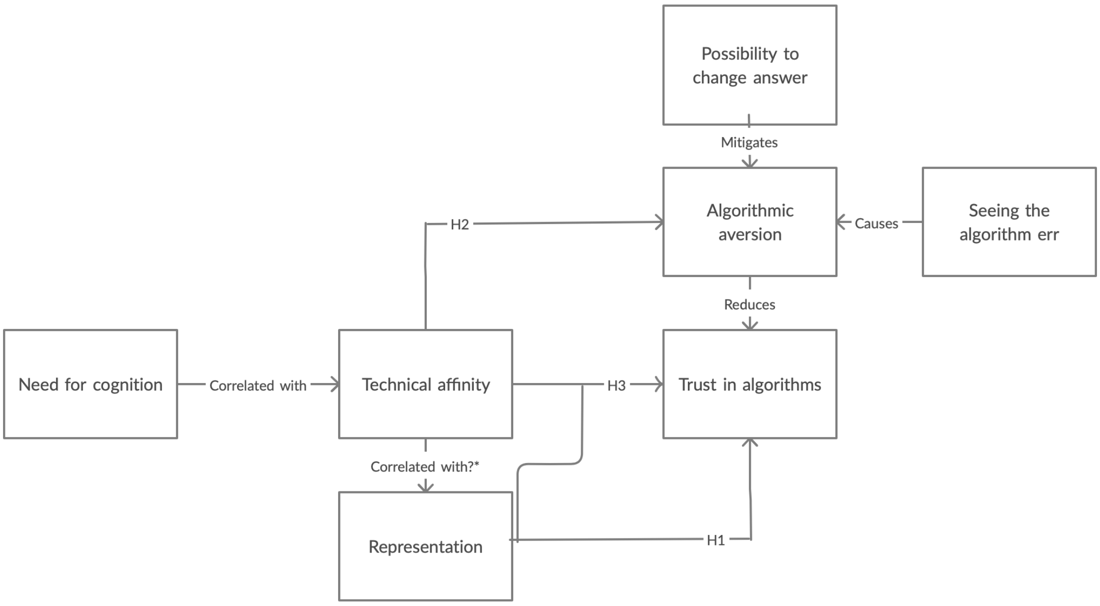


Figure 1: An overview of all the relevant variables found in previous work that affect the trust in algorithms. *This is the relation that is researched in H3 of this paper.

2.6 Key findings from previous work

A graphic overview of all variables influencing the trust of humans in algorithms can be seen in figure 1. Literature research into related work yielded the following key findings:

1. When confronted with a suggestion for an answer, people will to some extent conform to this answer, using it as anchor.
2. People place a lot of trust in algorithms, especially when it comes to objective, rational, and analytical tasks. This could be called algorithmic appreciation.
3. The algorithmic aversion effect: When people see an algorithm make mistakes, they lose trust in it while they do not lose trust in humans making the same mistakes.
4. Giving people the ability to change the predictions suggested by the algorithm reduces the algorithmic aversion effect.
5. People with low need for cognition (NFC) in general have low technical affinity.
6. People with a low NFC have more difficulties solving complex problems compared to people with high NFC.

3 Research focus

The main goal of these experiments is to try to find a connection between the representation of an algorithm and the trust placed in it. Secondly, data about technical affinity are collected to check if the effect is stronger or weaker for people with less affinity for technology.

While the experiments resemble the ones testing algorithmic aversion mentioned in the related work section, this is not the main focus of this study. Comparing the trust of the algorithm between the two phases might give results similar to the algorithmic aversion effect as of key finding 3. However, the study does not compare it to human advisors and key finding 4 suggests that giving people the possibility to change the algorithm's estimation mitigates the algorithmic aversion effect. Despite this finding, we still test it and if it is found we try to find a link between this effect and technical affinity.

The related work has shown that a lot of people tend to irrationally trust or mistrust algorithms. People with low NFC have difficulties solving complex tasks (finding 6) and might have more reason to offload the task to an algorithm. This might lead to them irrationally trusting an algorithm. People who have seen algorithms err more often (and thus probably have a higher ATI score) might have an irrational distrust of algorithms. The opposite could also be the case: they could have realistic expectations of algorithms due to their experience, be more open to using algorithms for various tasks, and/or be less susceptible to algorithmic aversion. Because the ATI test is based on NFC, it is interesting to see whether people with low ATI (and thus low NFC) trust the algorithm more than people with a higher ATI score and what its effect is on algorithmic aversion.

The hypotheses for this study are:

H1: People trust the outcome of algorithms more when they are presented with a representation of the algorithm.

H2: Both parts of the algorithm aversion effect are stronger for people with less technical affinity. They start with higher trust in the algorithm and lose more of this trust than people with more affinity for technology.

H3: The difference in trust between the representation and control group is larger for people with low technical affinity.

4 Method

This section outlines the methods used to conduct the study, including the tasks that are chosen, the different conditions and variables that are modelled to test the hypotheses and the overall procedure a participant undergoes while taking part in the experiments.

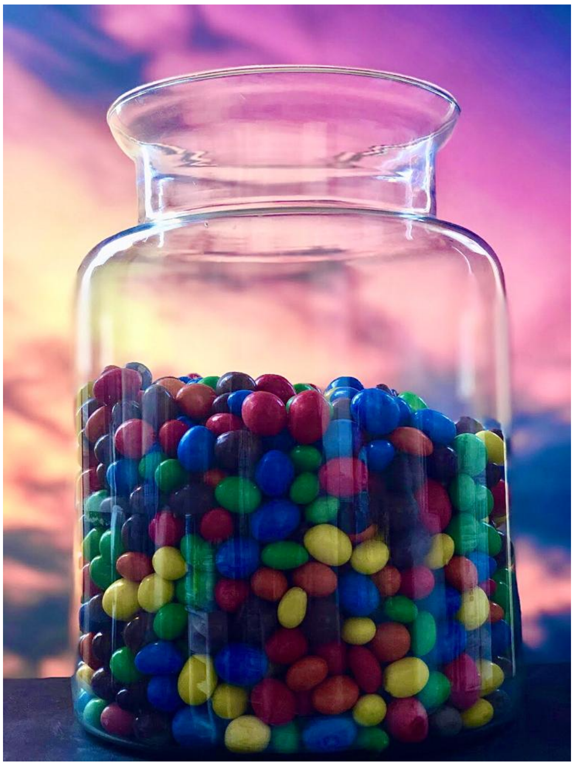


Figure 2: An example of the task participants had to perform. They had to estimate the number of M&M's in the vase.

4.1 Online vs offline experiment

The experiments are conducted online. for a description of the task in experiment II, see section 6.1. In experiment I, participants have to guess the number of M&M's in a glass vase based on the photos presented (for an example see figure 2). Offline (in-person) experiments were considered but abandoned in March 2020 when the situation regarding the COVID-19 pandemic made them impossible. During an offline experiment, the representation of the algorithm could be presented more vividly when the computer in the room is shown to actually perform the task (or convincingly fake it). The online variant relies on representation of an algorithm instead of demonstration. The representation described in the conditions and variables section below is used as main condition.

4.2 Task

This study requires tasks that are complex for humans, giving them an incentive to offload them to computers. We chose to use tasks that are not subjective, of social nature or have an emotional impact, since the related work showed that people might not want to use an algorithm for these tasks at all. This study is not about the ethical debate whether algorithms should be used for tasks that require social behaviour or empathy. We research the trust people have in algorithms that are used for analytical tasks. Finally, we wanted the tasks to be novel and/or fun. The participants recruited on MTurk fill out numerous surveys and therefore providing them with a novel task might incentivise them to pay more attention to it.

The task in experiment I to count M&Ms is complex because the volume has to be determined from a 2D picture. An algorithm has performed the same task and provides the participants with suggestions. Such an algorithm would consist of two parts: a computer vision element and an analytical element. The computer vision element extracts the number of M&M's from a picture or multiple pictures taken from different angles. The analytical part uses mathematics to determine the best guess of the number of M&M's based on the data.

Because this study focuses on the reaction of participants to the algorithm and considering the fact that building an algorithm to solve this problem is no trivial task, no real algorithm is used in experiment I and the answers given by the computer are predetermined. These estimations are presented to the participants and subsequently the participants enter their guess based on the information provided.

To test the algorithmic aversion effect there needs to be a temporal element to the experiments. This is a challenge because it should not be too obvious that the computer makes a mistake to test the reaction of the participants to this mistake. However, the participants cannot be overwhelmed by a stream of decisions since that would reduce the impact of each decision. To do this, the experiment consists of two phases of five estimations. In between these phases, the 'break' page shows the correct answers of phase 1 alongside the estimations made by the participant and computer. In the first phase, the trust of the participants is gained as they are presented with three fairly correct suggestions by the computer. However, two of the suggestions are outside the proposed accuracy boundaries.

During phase 2, the experiment continues with five more pictures to see to what extent people still conform to the computer's decision after they have seen the results of phase 1 and now know that the computer makes substantial mistakes.

4.3 Conditions and variables

In this section the conditions and variables that are used during the experiments are discussed. The experiments are performed under two conditions (representation and break page information) and the continuous variable technical affinity is measured and analysed.

4.3.1 Representation

To test the impact of how the algorithm is represented in the experiments, the participants are divided into two groups: the 'representation' group and the control group. Participants in the control group are not given any information about the workings of the algorithm while the representation group gains additional information before the experiments start.

We considered several methods to represent an algorithm while designing this experiment. When the experiments were still going to be conducted offline the focus was on confronting the participants with the system. Rotating the camera around the jar would give the feeling that the algorithm is actually scanning the contents. A (fake) algorithm running in a command prompt on the screen would suggest that the calculations are in progress. A possible extension to this setup would be to measure the difference in trust when the execution time of the algorithm is varied. These ideas could be replicated online by recording a video for each measurement, but this would not give the same experience as being in the room while it happens and seeing the algorithm perform 'live'. Therefore we abandoned this method when the experiments moved online.

Another idea was to give, for each task, an indication of the certainty that the algorithm performed within the given fault margin. The problem with this approach is that participants might answer in line with the algorithm when the probability is high and vice versa. This would lead to correlations between the certainty of the algorithm and the answers of the participants, but would not say anything about the trust in the algorithm.

The representation that we finally chose is to give one group more information about the algorithm and its workings. The representation group is told during the briefing that an advanced algorithm built by trustworthy developers is used in the experiment. They are also told that although the algorithm is advanced, it is not flawless and has an average error rate of about 12 percent. A short description of the algorithm explains that a rotating camera is used to get multiple angles of the jar that the algorithm bases its estimation on. The estimates are said to be rounded to the closest multiple of ten.

This form could lead to the problem that participants might suspect that they are dealing with a fake algorithm based on the description, because such an algorithm does not exist. The expectation is that just a small group of participants is able to deduce this from the information provided. The suspicious group probably also has a high score on technical affinity. Because the participants are asked to give a rating of the trust placed in the algorithm at the end of the test, these ratings can be correlated with the score for technical affinity to see whether this is indeed a problem. To test the trust of the group of participants with expert knowledge on computer vision algorithms, an extra question is added to let the participants assess their knowledge of computer vision algorithms.

4.3.2 Break page information

During the development of the experiments, we decided to add another small factor that resulted in four groups. During the break the results of the first round are shown. Adding the percentages that both the algorithm and the participant were off by alongside the absolute values gives more insight into the faults. It also enables comparisons to the accuracy mentioned in the extra information screen that is shown to the representation group. Although this provides valuable information, it makes the table rather big (see Appendix B for screenshots of both variants of the table). This could hinder the readability of this crucial information on a small-screen device, such as a smartphone, or a HiDPI monitor. Therefore the two extra columns are shown to half of the participants. With the aforementioned representation condition, this results in four groups. Our expectation is that there will not be a difference between the two break pages, since the participants can already calculate the percentages they and the algorithm were off by with the answers in the table. However, there is a chance that a clear overview of the percentages influences their opinion of the algorithm. If this second condition does not yield a significant difference, we merge the data for the two representation and the two control groups and use the original two groups for the analysis.

4.3.3 Technical affinity

The technical affinity of participants is a continuous variable in the analysis models. During the debriefing, participants are asked to fill out the ATI test (see Appendix A). The average of these questions results in an absolute score between 1 and 6 indicating technical affinity (where 1 is lowest and 6 is highest).

4.4 Procedure

The participants are randomly placed in one of the four groups and briefed on the estimation task and the fact that they are provided with algorithmic suggestions. To make sure that all participants are aware what an algorithm is, the following definition is mentioned (taken from similar research by Lee and colleagues.[16]):

> “Algorithms are processes or sets of rules that a computer follows in calculations or other problem-solving operations. In the situation below, an algorithm makes a decision autonomously without human intervention.”

The two groups that are presented with the representation page are given the information mentioned in the representation section and the other groups are led directly to the estimations.

When participants start the experiment, they go through the two phases of estimation with the break in between where the answers of phase 1 are shown. The participants are not limited by time to complete their estimations.

After the experiment we ask the participants to fill out a short questionnaire to get demographic information as well as their level of technical affinity, needed to test hypotheses H2 and H3. We also ask them to rate their trust in the algorithm on a Likert scale, as well as to give a self-assessment of their own performance and their knowledge of computer vision algorithms.

4.5 Amazon MTurk

To test the survey, we conducted a small pilot experiment where participants were recruited on social media channels. We used this pilot experiment to see whether the wording of the experiment could be improved and whether the data obtained were sufficient to run all proposed tests. We obtained the results of the main experiments using the Amazon MTurk platform. Access to the MTurk account of Snap Inc. was kindly provided by Maarten Bos.

MTurk provides an easy way to recruit large amounts of participants in a short period of time. Research has shown that the data quality is good, when certain precautions are taken[31][32]. Attention can be a problem and introducing attention checks improves the quality of data[33][34]. Therefore we added an attention check during the briefing of the representation group, where participants get information about the algorithm. To test their attention, people have to enter the mentioned accuracy of the algorithm (12 percent). The data of the participants who fail the check are excluded. This unfortunately only tests the attention of half the participants, but there is a lot more information on the briefing page for the representation group, making the need to test the attention of this group more important. We added a timer to the break page to make sure people spend enough time evaluating the results of the first phase.

We have taken some other precautions to increase the data quality. Only participants living in the United States are allowed to participate to help reduce a potential language barrier. Workers need to have completed at least 100 tasks and have an MTurk approval rating of over 90 percent. These are common MTurk requirements to improve the data quality of the responses.

4.6 Analysis

This section outlines all variables that result from the experiments and the statistical models and tests that are used in the result sections.

4.6.1 Variables

The experiments yield the following variables:

1. *Group*: A categorical variable defining whether the participants get to see the representation or the control briefing.
2. *Break page columns*: A categorical variable defining whether the participants get to see the two extra columns on the break page. The two extra columns show how much percent the algorithm and the participant were off by (see Appendix B for the two different break page tables).
3. *ATI*: The score for technical affinity resulting from the 9 questions in the ATI test. The range of this variable is between 1 and 6.
4. *Mean deviation in first half*: The mean of the absolute amounts the participant deviated from the algorithm's suggestion before the break in the experiments in which the intermediate results are shown.
5. *Mean deviation in second half*: The mean of the absolute amounts the participant deviated from the algorithm's suggestion after the break.
6. *Delta deviation*: The mean deviation in the second half minus the mean deviation in the first half of the experiments. Indicates the change in trust after seeing the intermediate results during the break and therefore shows the (presence or absence of the) algorithmic aversion effect. A higher delta deviation indicates an increase in deviation after the break and therefore a decrease in trust.
7. *Confidence*: The confidence of the participant in the performance of the algorithm, assessed after the experiments, using a five-point Likert scale.
8. *Self-assessment*: A self-assessment of participants' performance during the experiments, filled out after the experiments, using a five-point Likert scale.
9. *Computer Vision*: Participants' self-assessed familiarity with computer vision algorithms, assessed after the experiments, using a five-point Likert scale.
10. *Age*: Participants' self-reported age, used as a continuous variable.
11. *Gender*: The possible answers are "male", "female", "other", and "prefer not to say".

4.6.2 Algorithmic aversion

To test whether the algorithmic aversion effect occurred during the experiments, we compare the mean of the absolute deviations *before* and *after* using a two-sided paired *t-test*. We do this with the complete data set and separately for the participants who saw the representation page and those who did not.

If the effect indeed occurs, we compare the difference in *delta deviation* between the representation and control groups to see if there is a difference in effect between the two groups. We use this difference to get a first impression of the answers to the hypotheses.

4.6.3 Building models

After getting a first impression of the algorithmic aversion effect, we build a model to check the significance of all factors affecting the trust. Deviation *before*, *after* and the *delta deviation* are the response variables in the models with group, ATI score, age, and gender as predictors. Even if the algorithmic aversion effect is not found, the trust before and/or after the break can be significantly different.

From the background research there is reason to believe that the means of representation have an influence on the result[21]. Out of our own curiosity, we added the technical affinity condition to the model, to see if it has an impact on the results. Therefore we built three models for the three response variables, with these two variables as predictors. We compare these to models which include age and gender to check if they also have a significant effect, although this is not expected.

We use a Tukey post-hoc test on the group predictor variable to conclude whether the group participants are in has an influence on the trust in the algorithm (H1).

We check the significance of the ATI variable's effect on the three different models. By comparing the deviations *before* and *after* we examine whether there is a difference in trust before and after the break. If both parts of the algorithmic aversion effect are indeed stronger for people with less technical affinity (H2), the deviation *before* should be smaller and the *delta deviation* should be larger for people with low technical affinity.

The models contain the interaction between technical affinity and the group variable. We analyse this interaction to see if the difference in trust is larger for people with low technical affinity (H3). After checking this effect for all the data, we remove the entries for which the ATI score lies one standard deviation above or below the mean from the data set, only leaving the data for participants with either a low or high ATI score. We model and test the resulting values in the same way as before.

4.6.4 Self-assessment

We use the confidence variable to check whether the use of deviation from the suggestion made by the algorithm is indeed a good measure of trust placed in it. If it is a good measure, the self-assessed trust in the algorithm should correlate with the deviation. We use a Pearson correlation to check this.

Similarly, we use the confidence variable to check if the description of the algorithm in the representation group leads to suspicion that the algorithm is fake. If people with high technical affinity gave a very low score on confidence, they probably saw through the fake algorithm. We use a Pearson correlation to test this.

4.6.5 Task difficulty

Our prediction is that the task is quite hard for the participants. We use a repeated measures analysis to check whether the performance of the participants improves over the course of the ten tasks. Since feedback is provided half-way through the experiment, it is very likely that the performance in the second half is better than in the first.

5 Experiment I: M&Ms in a vase

In this section we present the results from both the pilot as well as the final experiment. We used the R programming language for all statistical tests in this section. Significance for all these tests is determined by $\alpha(0.05)$.

		Mean deviation before		Mean deviation after		ATI	
Group	N	Mean	SD	Mean	SD	Mean	SD
Control	12	235	139	143	77.8	3.70	0.853
Control with extra column	11	298	277	221	142	4.06	0.682
Representation	10	220	112	178	61.3	3.59	0.918
Representation with extra column	9	173	105	114	54.0	3.75	0.853

Table 1: The most relevant information for the four groups in the pilot experiment ($N = 43$).

5.1 Pilot

We use the data set obtained in the pilot phase ($N = 43$, $M_{age} = 33$, for the most relevant data per group see table 1) to check whether the participants understood the survey or needed updates for clarification. It is also a test case for the proposed statistical tests in the final experiment.

The comparison of mean deviation before and after the break showed that participants followed the suggestions of the algorithm significantly more after the break ($t(43) = 2.26$, $p = 0.029$, $d = 0.706$, mean difference = 69). This is the exact opposite of the expected result based on the algorithmic aversion effect. It could be that people get better at the task at hand after a few pictures and therefore do not follow the algorithm more but just have a better idea how many M&M's are in the vase. Another explanation is that the errors made by the algorithm during the first part of the experiment are not severe enough to cause doubt about its performance. If this is the case, the occasions where the algorithm is within the promised accuracy range could have helped build confidence in the algorithm, resulting in better trust in the second part.

When evaluating the groups individually, the significant effects disappear. This could indicate that the sample size is not large enough to narrow the data down to multiple groups. The pilot data do not contain significant interactions between the group variable and the deviations.

No relevant implications for the ATI factor were found. The p-value of the ATI factor was far above 0.05 in all three models. All other subsequent tasks performed did not yield significant data. The pilot was useful to check the survey, but for substantial results more participants are needed.

Based on the feedback, some minor details were changed for the full experiment. The pictures did not entirely fit on the screen of a HiDPI display, therefore we reduced the pictures slightly in size. They now fit on the screen but still are large enough to get a good overview of the task. We made the answer suggested by the algorithm that was least accurate even less accurate to accentuate the major mistake of the algorithm during the break. We reduced the time people had to wait at the 'break' page from 20 seconds to just 5 seconds. The pilot showed that people paid enough attention to the presented table to reduce the time constraint which could 'lock' them into the experiment. We added a timer to the break page to filter out participants that clearly did not spend enough time on the page to read all required information. Apart from these changes there were only a few minor textual adjustments.

5.2 Results

In this section we analyse the results of experiment I and discuss the effect of the conditions and variables on the deviation from the algorithm's suggestions. We also introduce some post-hoc analyses based on the time spent on the break page and the performances of the participants and algorithm.

5.2.1 Data

For the final experiment we recruited 300 participants on Amazon MTurk, who were randomly assigned to one of the four groups. After filtering out the people that did not pass the aforementioned attention checks, a data set of $N = 284$ remained ($M_{age} = 38$ years; 190 male, 91 female, 2 other and 1 preferred not to disclose gender; 152 in the control groups and 132 in the representation groups). For the most relevant data obtained during the experiment, see table 2. During the analysis we realised that the correct answers to the tasks after the break were quite a lot higher than

		Mean deviation before		Mean deviation after		ATI	
Group	*N*	Median	SD	Median	SD	Median	SD
Control	78	26.5	36.5	19.2	12.9	4.03	0.747
Control with extra column	74	20.2	18.9	18.0	14.6	4.02	0.871
Representation	63	13.8	13.6	17.8	16.1	4.18	0.771
Representation with extra column	69	17.9	17.6	17.9	15.4	4.07	0.835

Table 2: The most relevant information for the four groups in the experiment.

		Mean deviation before		Mean deviation after		ATI	
Group	*N*	Median	SD	Median	SD	Median	SD
Control	152	23.4	29.4	18.6	13.7	4.03	0.808
Representation	132	16.0	15.9	17.8	15.6	4.12	0.804

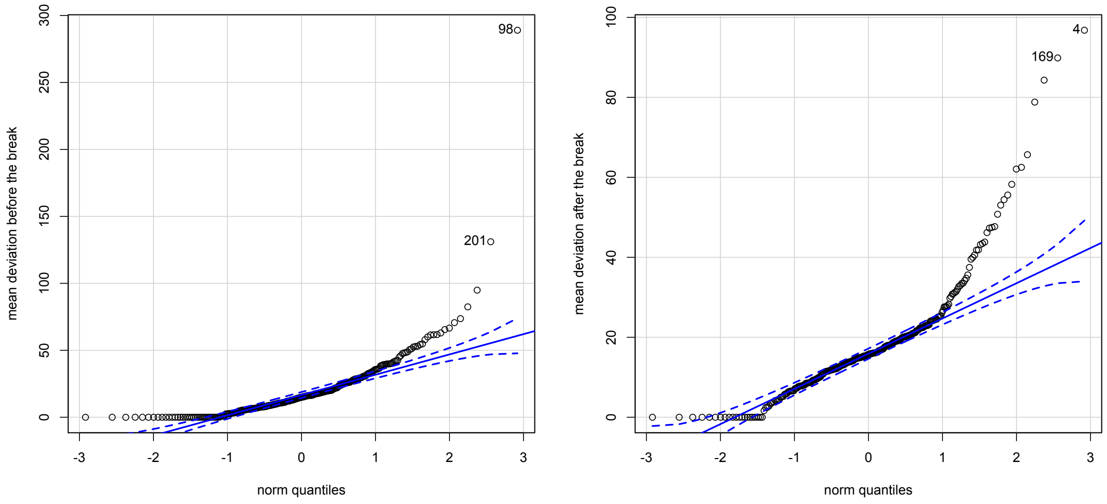
Table 3: The most relevant information for the two groups in the experiment after we lifted the "break page time" condition and merged the two control groups and the two representation groups.

the ones before the break. To compensate for this, we decided to use the the absolute percentages participants deviated from the suggestions instead of the absolute deviations.

As mentioned, we reduced the time constraint on the break page from 20 seconds to 5 seconds based on the outcomes of the pilot. An added timer still gave us the opportunity to exclude participants based on time spent on the break page, which is the crucial point where the change in trust should occur if the algorithmic aversion effect holds. Excluding participants based on the time they spent on this page substantially changed the results and therefore both the original data as well as data sets where the break page visit time is greater or equal than 10 and 20 seconds were evaluated.

We reduced the four groups, as mentioned in the method section, to two, because the extra columns in the table on the break page did not have an impact on the results. The Tukey post-hoc test did not yield significant differences between the groups that were shown the extra column and their counterpart without the extra column. We only found significant differences between the representation and control groups without the extra columns. The models for the deviation before ($p = 0.011$ and a mean difference of 13) and delta deviation ($p = 0.027$ and a mean difference of 11) showed significant differences for these two groups. Table 3 shows the data resulting from removing the "break page information" condition. This data set is used for the rest of the analysis.

The deviations from the suggestion made by the algorithm turn out to be far from normally distributed (see figure 3). Therefore we decided to use nonparametric tests where these equivalents were available to compare the medians instead of the means. We replaced the t-tests by Mann-Whitney-Wilcoxon tests and the proposed Pearson correlations by Spearman correlations. We still used the ANCOVA tests because they were not easily replaced by a nonparametric variant. We tried to find conclusive proof whether ANCOVA tests are robust against non-normality in large sample sizes. We found arguments for both standpoints and finally decided to use the ANCOVA models as they are the best option for our data.



(a) Mean of the (percentage of) deviations before the break.

(b) Mean of the (percentage of) deviations after the break.

Figure 3: Q-Q plots to check the mean estimations for normality. The line representing the model clearly is not a good fit for the outliers. The median is taken from the mean deviations before and after the break. The mean deviation is the average percentage a participant deviated from the algorithm.

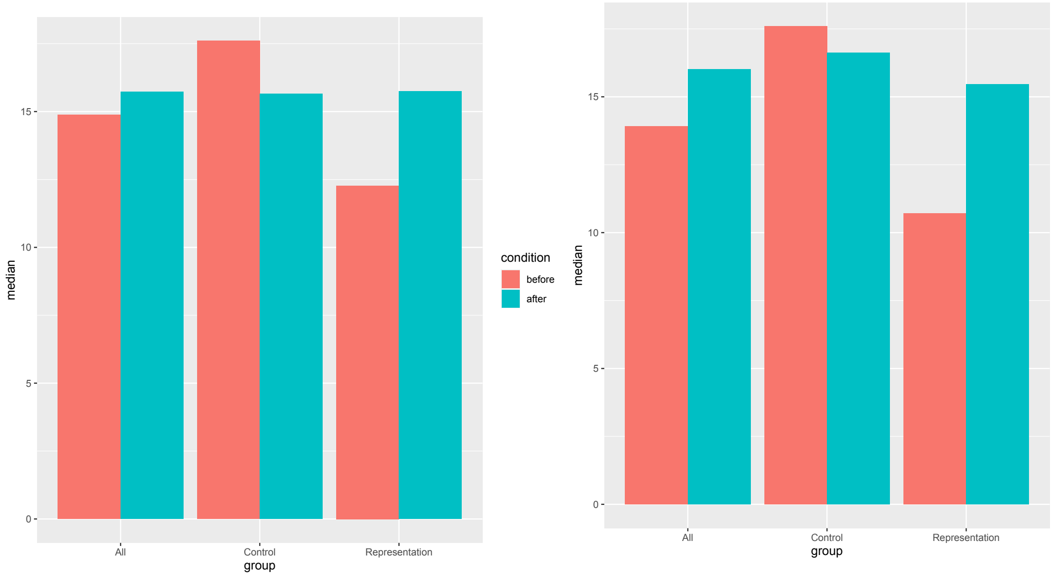
5.2.2 Groups

We calculated the percentages people deviated from the algorithm's suggestion for each task. For each participant, the mean deviation before is the average of these percentages for tasks 1 to 5. Similarly, The mean deviation after is the average for tasks 6 to 10 per participant. The medians of these outcomes are plotted in figure 4a for all data, the control group and the representation group respectively.

We tested the significance of the changes with paired Mann-Whitney-Wilcoxon tests, comparing the medians before and after the break. It turned out to be almost significant for the representation group (median before = 12.3, median after = 15.7, Z = -1.82, p = 0.07 and r = -0.42), but not significant for all data and the control group (all data: p = 0.96 and control: p = 0.13). Taking the data with a break time of at least 10 seconds instead of five (N = 240, 132 in control, and 109 in representation), the medians change only slightly, but enough to make the change in the representation group significant (median before = 12.3, median after = 15.8, Z = -2.1, p = 0.038 and r = -0.70). Evaluating only the data with a break time of at least 20 seconds (N = 154, 83 in control and 71 in representation) changes the results (see figure 4b). The difference between the two phases for the representation group is now very profound (median before = 10.7, median after = 15.5, Z = -2.5, p = 0.014 and r = nan). The calculation of the r value returns with an error that it is not a number. We have not been able to establish why this error occurs. The significance values did not change for all data and the control group while varying the break page time.

The three ANCOVA models with *before*, *after* and *delta deviation* as response variables, group as independent variable and ATI as covariate confirm the aforementioned results. The models with *before* and *delta deviation* as response variables show a significant effect of the independent group variable (*before*: $F(1,\ 280) = 6.8$, $p = 0.001$, $d = 0.31$ and *delta deviation*: $F(1,\ 280) = 5.7$, $p = 0.018$, $d = 0.28$), while in the model with the *after* response variable the group variable is not significant ($p = 0.65$). The Tukey post-hoc tests on the comparison of means within these models confirm the found effects. The same p-values are found and the tests show the differences in medians between the groups for the *before* (7.46, higher in the control group) and *delta deviation* (6.68, higher in the representation group) response variables. The data for break times of five and 10 seconds show similar significant values for the group variable.

People in the representation group spent more time on the break page deviated more from the algorithm, showing a big drop after the break. This drop happens while the control group actually relies on the algorithm more. Either the control group gains confidence in the algorithm after seeing its results or they get better at the task. If people indeed get better at the task, it is even more interesting that despite this the deviation increases for the representation group. There is reason to believe that people indeed get better at the task. After the first five pictures people get an impression of how many M&M's fit in the vase. Based on this assumption they can evaluate how full the vase approximately is and calculate their estimation accordingly.



(a) Median of percentual deviations for all data (N = 284).

(b) Median of percentual deviations for participants who spent at least 20 seconds on the break page (N = 154).

Figure 4: Median of the mean percentual deviations from the algorithm's suggestions per participant before and after the break, for the two groups and all data.

The repeated measures analysis did not turn out to be useful for our experiment. We planned to run the analysis on the 10 tasks, but realised that there is only one moment in the experiment where the circumstances of it change (the break page). Therefore the significant differences between individual tasks before and after the break do not provide extra information that cannot be gathered by the ANOVA models.

5.2.3 ATI

The distribution of the ATI variable can be seen in figure 5, where 1 is low and 6 is high. The ATI values are on the higher end of the spectrum (mean = 4.03). This could be due to the fact that we all use technical devices in our daily lives and have therefore developed a baseline of technical affinity, or we at least believe we do (see Dunning-Kruger effect). The values seem normally distributed, but the Shapiro-Wilk test for normality shows they are not ($W = 0.98$ and $p = 0.00031$). Although the ATI values are on the high end of the spectrum, there is quite some variance which allows for variance testing using ANOVA models.

The models do not show evidence that technical affinity has an influence on the response variables in the three models (before: $p = 0.28$, after: $p = 0.82$ and delta: $p = 0.33$). Emphasising the higher and lower values by removing all participants with an ATI score one SD below and above the mean ATI results in a data set with $N = 81$, 40 of which are in the control group and 41 in the representation group, $M_{ATI} = 4.21$, SD = 1.37. Using the same models on this data set decreased the probability values for the influence of ATI on the response variables, but they are still far from significant (before: $p = 0.36$, after: $p = 0.57$ and delta: $p = 0.21$). Because this effect does not even occur in a data set with only the high and low values, we can safely assume that there is no effect of the measured technical affinity on the trust in the algorithm used in this experiment.

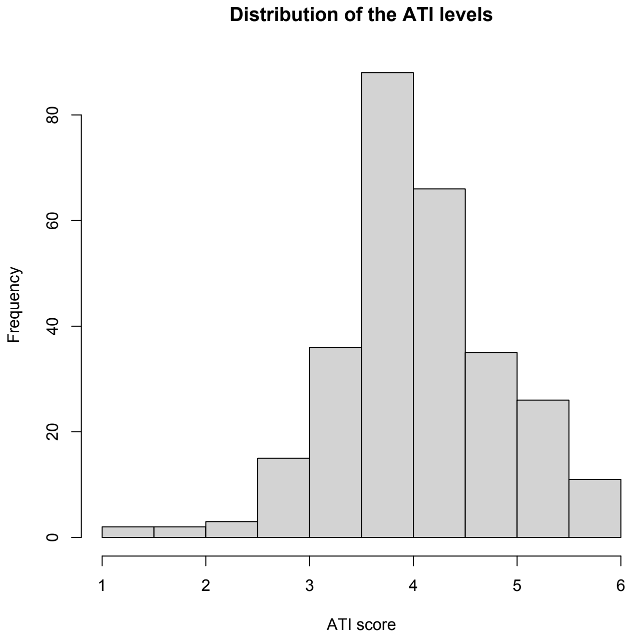


Figure 5: Plot of the ATI scores of the participants. Clearly the values are on the high end of the spectrum.

5.2.4 Trust in the algorithm

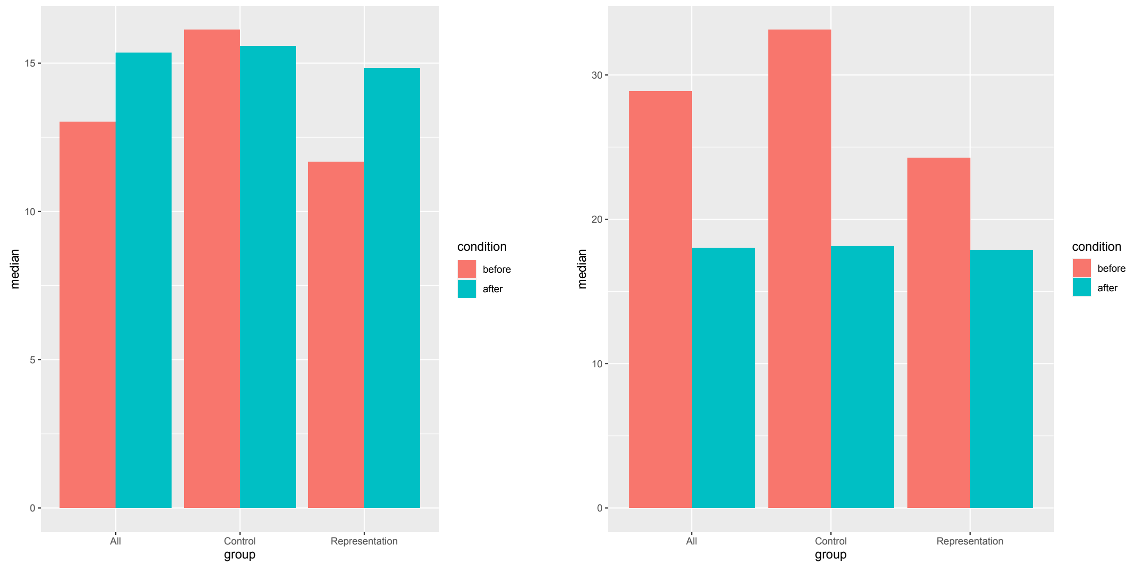
We added some self-assessments at the end of the survey test to evaluate whether the *delta deviation* is a good measure for the confidence in the algorithm. After seeing the non-normality of the data (see section 5.2.1 and figure 3), we replaced the planned Pearson correlations by Spearman correlations. A Spearman correlation between the *delta deviation* and the self-assessed confidence in the algorithm yields an almost significant negative correlation ($rs(282)$ = -0.10, p = 0.087). This changed when we evaluated only the data for the people who spent at least 10 seconds on the break page ($rs(238)$ = -0.14, p = 0.033). It could be that the proposed test is not suitable for evaluating the *delta deviation.* The confidence is only recorded after the experiment, while the *delta deviation* is a measure of the confidence throughout the whole experiment.

In an attempt to asses whether people with expert knowledge would notice that the algorithm was fake, we asked participants to rate their familiarity with computer vision (CV) algorithms. Only very few participants are self-reported experts in the CV field (mean CV score: 2.88/5, only 26 participants rated their familiarity as 5/5). Both the CV assessment of the participants as well as their ATI score are correlated with the self-assessed trust in the algorithm. The CV score showed a very clear positive correlation with the confidence ($rs(282)$ = 0.58, p < 0.001), while the ATI score does not correlate significantly ($rs(282)$ = 0.016, p = 0.79). However, removing participants with an ATI score one SD above and below the mean to accentuate the high and low ATI scores does yield a significant positive correlation ($rs(79)$ = 0.22, p = 0.048).

From the positive correlations we can assume that there is no (extra) doubt about the realness of the algorithm by experts. This is unexpected, as the description of the algorithm is rather vague and could provoke quite some questions. The positive correlation between CV and confidence is even quite strong. This indicates that participants with expert knowledge actually had more confidence in the algorithm compared to laymen.

5.2.5 Performance

After suggestions made during the presentation of a draft version of this thesis, we decided to compare the participants that performed better than the algorithm in the first phase of the experiment to the ones that performed worse. This was done to see how both groups react after they have compared their performance to the algorithm's during the break.



(a) Median of percentual deviations for participants who performed better than the algorithm during the first half. (N = 135).

(b) Median of percentual deviations for participants who performed worse than the algorithm during the first half. (N = 111).

Figure 6: Median of the mean percentual deviations from the algorithm's suggestions per participant before and after the break, for the two groups and all data.

Figure 6 shows the median percentual deviation from the suggested answers for the group that scored better than the algorithm (left in the figure) during the first half of the experiment and the group that scored worse (right in the figure). The group that conformed to the algorithm completely and therefore scored equally was quite small (N = 38) and was excluded from this analysis.

The median *delta deviation* for the participants that scored better than the algorithm shows significant loss of trust after the break (median before = 13.0, median after = 15.3, Z = -3.14, p = 0.0017 and r = -0.70). This is caused by the participants in the representation group, who lose a lot of trust after they perform better than the algorithm (median before = 11.7, median after = 14.8, Z = -3.17, p = 0.0015 and r = -0.70). However, the control group conform more to the algorithm after the break even when it performs worse than themselves (median before = 16.1, median after = 15.6, p = 0.23). The representation group got promised that the algorithm has a mean error of 12 percent. Therefore they seem to start with significantly more confidence in the algorithm than the control group according to the data (median control before = 16.1, median representation before = 11.7, Z = 1.86, p = 0.025 and r = 0). The groups deviate similarly from the algorithm in the second half of the experiment. The influence of the group condition on the *after* and *delta deviation* response variables is not significant (both $p > 0.38$) and the difference in median *delta deviation* is also not significant (p = 0.52). This means that although there seemingly is a big difference between the groups, this is caused by the median values of the variables and the individual pairings of *before* and *after* do not reflect this same difference.

When the participants in the representation group see that they perform better than the algorithm and that in 2 out of 5 of the tasks before the break the expected fault margin is not met, they deviate significantly more. The control group starts without this promise and sees that the algorithm performs quite well at 3 out of 5 tasks. Even when they are better than the algorithm, they conform more after the break. They probably see it as a useful tool to anchor their own suggestions (more on this in section 5.2.6). Although the drop in trust for all the data is in line with the algorithmic aversion effect, this is due to the drop in the representation group. This drop can therefore not be explained using the algorithmic aversion effect, but is probably due to the algorithm not delivering the results that were promised to the representation group.

The participants who scored worse than the algorithm adhered to its suggestions more after the break on average, regardless of the group they were in. All these drops in *delta deviation* were

Group	N	Mean performance before Median	SD	Mean performance after Median	SD
Algorithm	1	26.3	NA	12.8	NA
Better - All	135	22.4	4.63	16.2	10.2
Better - Control	65	21.8	4.41	16.8	10.6
Better - Representation	70	23.2	4.82	15.4	9.78
Worse - All	111	33.6	21.6	20.8	15.5
Worse - Control	67	35.7	25.0	19.4	13.9
Worse - Representation	44	31.7	15.1	21.2	17.8

Table 4: The performances of the participants performing better and worse than the algorithm, per group (lower scores are better). The mean performances are the means of the percentual deviations from the right answers per participant.

significant (all: median before = 28.9, median after = 18, Z = 4.71, p < 0.001 and r = -0.31; control: median before = 33.1, median after = 18.1, Z = 4.27,p < 0.001 and r = 0; representation: (median before = 24.2, median after = 17.8, Z = 2.11, p = 0.03 and r = 0). Similar to the participants that performed better than the algorithm, the people in the representation group start with more confidence in the algorithm than people in the control group, although now the difference in *delta deviation* between the groups is not significant (median control = 33.1, median representation = 24.2, Z = -1.81, p = 0.64 and r = 0). The influence of the group condition on all three response variables is not significant (before: p = 0.12, after: p = 0.70, delta deviation: p = 0.07).

An explanation for the non-significant effect of the group variable on the results could be that the worse performing representation group is conflicted. In this situation part of the group conforms more to the algorithm because it performed better than them, while others deviate more from it because their expectations are not met. If this reasoning is correct, they cancel each other out to create a smaller effect on the medians of the worse performing representation group.

Looking at the performances for the participants that scored better or worse than the algorithm (see figure 4), the group that performed worse than the algorithm improves significantly in the second half of the experiment by trusting the algorithm more (Z = 7.54, p < 0.001 and r = 0.71). This is partly due to the fact that the algorithm performs a lot better in the second half of the experiment, but it is still a substantial improvement. The scores of the group that performed better than the algorithm in the first half drop after they deviate more from it during the second half (Z = 4.57, p < 0.001 and r = -0.71).

The representation group also improves their performance in the second half. This seems contradictory when the algorithm performs better in the second half and it thus seems a good idea to follow its advice. It could be an indication that the participants in this group deviate from the algorithm more when it errs substantially, but conform to it when it performs well. This would indicate good use of the algorithm as anchoring tool, while still using common sense to weed out its mistakes. This assumption is evaluated in the following section.

5.2.6 Using the algorithm as an anchoring tool

To evaluate whether the participants used the algorithm as anchoring tool to adjust their own guesses in the second phase, we calculated the deviations from the *good suggestions* (those that were within the promised 12 percent error margin) and the *bad suggestions* separately for each participant. There were three good suggestions and two bad suggestions in both the first and second phase.

There is a significant difference between the median deviations from the good suggestions and the bad suggestions of the algorithm (median good = 12.2, median bad = 19.6, Z = -7.33, p < 0.001 and r = -0.364). This indicates that on average participants use the tool wisely: they conform to it more when it seems to be on the right track, but keep aware that it can err.

This difference in deviations leads to a significant difference in performance (median good = 13.7, median bad = 24.3, Z = -9.69, p < 0.001 and r = -0.704). The participants perform a lot better on the tasks where the algorithm provides good suggestions. In the second half the algorithm performs quite well (see table 4), and therefore conforming leads to good performance.

5.3 Discussion

This section discusses the results we found in the experiment, the mistakes we made as well as the limitations of it. Finally, we discuss the improvements we made for experiment II.

The significance of the group factor in the ANCOVA models as well as the subsequent significant Tukey post-hoc test show that there is indeed a difference in the trust in algorithms between the two groups. This difference only occurs before the break page, where the representation group has more confidence in the algorithm in line with H1. After the break there is a decline in trust in the representation group and an increase in trust in the control group, bringing them in line with each other. H1 can therefore not be accepted, as the difference in trust only occurs in the first half of the experiment.

We did not significantly reproduce the algorithmic aversion effect for all data. In general, people did deviate more from the suggestions after the break, but this change was not significant. The fact that it is not significant is due to the control group which actually gets closer to the suggestions. An explanation could be that the effect is mitigated by the fact that people can change the suggestions to their own amount instead of an accept/reject decision. Previous research showed that this increases trust in algorithms[20]. It could also be that people were more accustomed to the task in the second phase and therefore scored better.

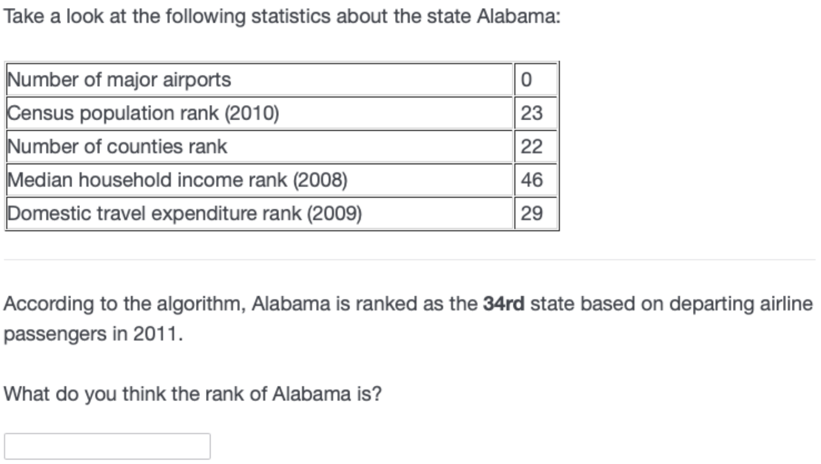
If people indeed get better at the task, it makes the results for the representation group even more remarkable. There is already quite a big decrease in trust in this group after the break. If people get better at the task, they would deviate from the algorithm less. If the decrease in trust is mitigated by the fact that people get better at the task, it should be even more noticeable for different tasks.

The data show no significant effect of the ATI scores on the algorithmic aversion effect or the group differences. On this basis, H2 and H3 are not accepted. This is not that surprising, as there was no precedent for this effect. It was purely our own curiosity that drove us to test this relation. There are several factors that complicate research into this topic. The Dunning-Kruger effect affects people's assessment of their own capabilities. While the authors of the paper on the ATI test try to circumvent this by averaging nine questions to make the connection between the questions and the effect measured not too obvious, there were almost no participants with a rating of one or two out of 6. This could also be due to the fact that everyone uses technology in their daily lives and has overcome the worst struggles with it. While these are realistic scenarios, it may also very well be possible that the test is accurate and there just is no effect of technical affinity on the trust in algorithms.

There is a positive correlation between the self-assessed confidence in the algorithm and the ATI score of participants. From that relation we can conclude that if there is a difference in trust between people based on technical affinity, contrary to the hypothesis people with high ATI scores have more confidence in the performance of the algorithm. The assessments were only done after the experience and therefore we can't use these data to asses the change in trust (H2).

The analysis of the participants' performance showed that conforming to the algorithm increased the score during the second half of the experiment. People performing worse than the algorithm during the first half improved a lot by deviating less from it in the second half. The difference in deviation from the algorithm's bad suggestions and good suggestions indicates that people use the suggestions as an anchor but still use their own judgement to spot blatant mistakes by the algorithm.

The speed at which responses are gathered using Amazon MTurk enabled us to run an extra experiment to test whether the specifics of the task in the first experiment and the errors made while conducting that experiment had an influence on the results. After seeing the results, there was some reason to believe that people get better at the task because they approximately know how many M&Ms fit in the vase after they have seen the correct answer to a question where the vase is quite full. The imbalance between the answers in the first and second phase could also have influenced the results. In the following experiment the algorithm errs exactly the same amount



Take a look at the following statistics about the state Alabama:

Number of major airports	0
Census population rank (2010)	23
Number of counties rank	22
Median household income rank (2008)	46
Domestic travel expenditure rank (2009)	29

According to the algorithm, Alabama is ranked as the **34rd** state based on departing airline passengers in 2011.

What do you think the rank of Alabama is?

Figure 7: An example of the task in the "Airline passengers" experiment.

in the first and second phase. Finally, the briefing of the representation group in the second experiment does not mention that it is developed by a respectable company. This could have given the algorithm authority which may have led to the larger amount of trust in the representation group before they see it err.

6 Experiment II: Airline passengers

For the second experiment of this study, we use almost the same methods as described in the method section for this paper. However, we made some adjustments after experiment I. We lifted the extra condition where half the participants saw two extra columns on the break information page as it did not yield different results. Therefore there are only two groups in this second experiment: the representation group and the control group.

We used the exact same survey and analyses for the second experiment, the only things we changed in the survey were the task and the representation text page.

6.1 Task

To contrast the visual task in the "M&Ms in a vase" experiment, it was replaced by a text-based task where the participant has to guess the rank of a state in the United States based on its number of departing airline passengers in 2011. This is the same task that Dietvorst used in his 2015 paper and he kindly provided us with the supplementary resources needed to use the task in this experiment [7]. To help with this task, the participants are provided with some statistics about the state (for an example of the task see figure 7). Like in the previous example, they get a suggested answer calculated by an algorithm.

Contrary to the previous experiment, we provided the participants with suggestions that Dietvorst calculated using a real algorithm. Dietvorst calculated the predictions for these ranks using an ordered logistic regression and the average fault of the algorithm is 4.32 ranks. The information that an ordered logistic regression was used and the average fault are communicated to the representation group before the experiment. The control group only learns that an algorithm has calculated the suggestions, but not how it did it and what its accuracy was.

6.2 Results

For this experiment, we again recruited 300 participants using Amazon MTurk. Of these participants, $N = 272$ ($M_{age} = 37$ years; 184 male, 88 female; 145 in the control group and 127 in the representation group) remained after ruling out the people who failed the attention checks or misunderstood the task. A few participants ($N = 9$) filled in very high values for the ranks. Since the United States only has 50 states, all submission with answers higher than 60 are discarded.

		Mean deviation before		Mean deviation after		ATI	
Group	N	Median	SD	Median	SD	Median	SD
Control	145	2.6	5.2	3.2	5.9	4	0.71
Representation	127	1.2	4.1	2.8	4.1	4	0.79

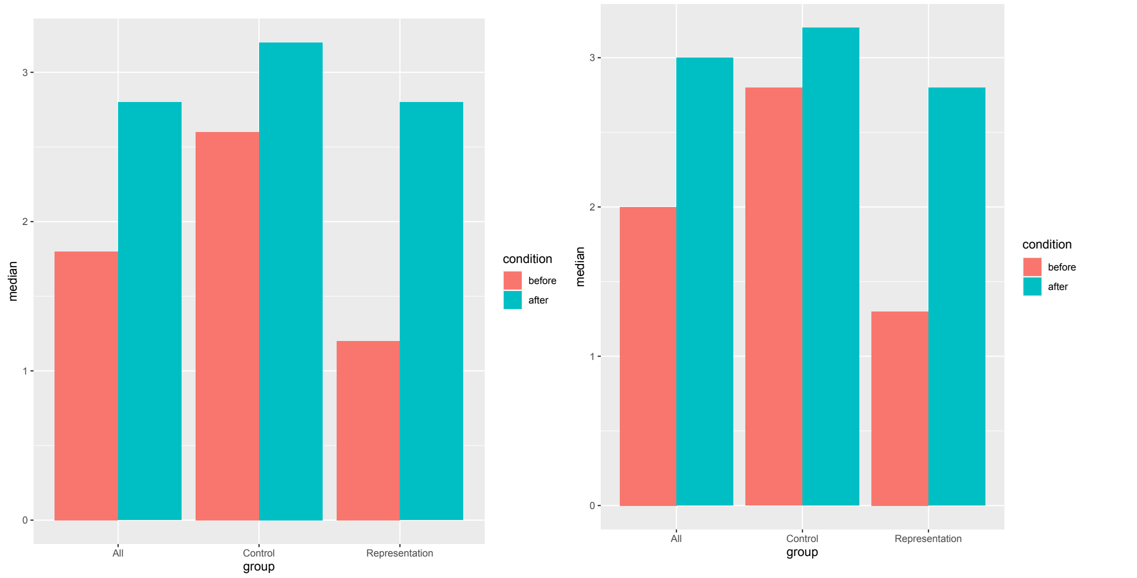
Table 5: The most relevant information for the two groups in the “Airline Passengers” experiment ($N = 272$). The medians are absolute values.

This leaves some room for error for people who do not know how many states the United States consists of. Since all participants recruited are from the US, it is safe to assume that values above 60 are caused by different misunderstandings.

Because the differences between the answers in this task are not as big as in the “M&Ms in a vase” experiment (they vary between 1 and 50), we use the absolute guessed ranks instead of the percentual guesses in the previous experiment. The most relevant data can be seen in figure 5.

6.2.1 Groups

The differences between the deviations before and after the break are significant for both groups individually and for all data combined (all data: median before = 1.8, median after = 2.8, Z = -4.85, $p < 0.001$ and r = -0.44; control: median before = 2.6, median after = 3.2, Z = -2.63, p = 0.0085 and r = -0.52; representation: median before = 1.2, median after = 2.8, Z = -4.32, $p <$ 0.001 and r = -0.38). All three deviate a lot more after the break page, losing significant trust in it (see figure 8a). This is in line with the algorithmic aversion effect and could be an indication that the doubts in the first experiment (that the effect was mitigated by the fact that people get better at the task) are justified.



(a) Median of absolute deviations for all data (N = 254).

(b) Median of absolute deviations for participants who spent at least 20 seconds on the break page (N = 155).

Figure 8: Median of the mean absolute deviations from the algorithm’s suggestions per participant before and after the break, for the two groups and all data.

The plot of the data set with only participants who spent at least 20 seconds on the break page (N = 155) shows a smaller drop in trust for the control group, which is now not significant anymore (median before: 2.8, median after: 3.2, p = 0.09). The representation group and the combination of the two groups still show a significant drop in trust (all data: median before = 2, median after =

3, Z = -3.66,p < 0.001 and r = -0.39; representation: median before = 1.3, median after = 2.8, Z = -3.75,p < 0.001 and r = -0.25). This could mean that people who spent more time on the page realised that the algorithm was quite accurate in the majority of tasks. In that case participants in the control group should conform to the algorithm more with reasonable suggestions. This is investigated further in section 6.2.4. The participants in the representation group are clearly disappointed in the algorithm for performing worse than the promised average error.

The three ANCOVA models with *before*, *after* and *delta deviation* as response variables, group as independent variable and ATI as covariate show a significant effect of the group condition on the deviations before and after (*before*: $F(1, 268) = 11$, $p < 0.001$, $d = 0.42$ and *after*: $F(1, 268) = 6.8$, $p = 0.010$, $d = 0.33$). The model with the *delta deviation* as response variable strangely does not show a significant effect of the group variable ($p = 0.50$). Looking at the plot (figure 8a) there seems to be a big difference in deviation between the two groups (control: 0.6, representation: 1.6). Therefore we performed a Kruskal-Wallis rank sum test to check whether this result could have been caused by the non-normality of the deviations, but this test is also not significant ($H(1)$= 1.67, $p = 0.20$). Apparently the individual *delta deviations* of participants combined are not significantly influenced by the group condition, although the medians of the *after* and *before* suggest there is a difference.

The ATI does not have an influence on the response variable in the three models (all $p > 0.30$). Even accentuating the higher and lower ATI scores by removing all participants one standard deviation above and below the mean ATI score did not reveal any indication that ATI has an influence on the response variables. After finding no significant results in the two experiments, the conclusion is that there either is no effect or the test used is not a good indicator for technical affinity.

6.2.2 Trust in the algorithm

To check whether the *delta deviation* is a good indicator for trust in the algorithm, we correlated it with the participants' self-reported confidence in the algorithm. This correlation is not significant ($p = 0.95$). People who deviate less from the algorithm thus do not have significantly more confidence in it. However, they might use it as an anchoring tool and trust in a combination of the algorithm's suggestion and their own judgement combined. More on this in section 6.2.4.

There is a positive correlation between the familiarity of the participants with regression models and the self-reported confidence in the algorithm ($rs(270) = 0.25$, $p < 0.001$). We tested this to see whether there was doubt about the realness of the algorithm with experts. While the algorithm used is real in contrast to the fake algorithm used in experiment I, participants do not see it in action and could still think the numbers are made up. The positive correlation could imply that being familiar with regression models gives the right expectations of its performance.

The ATI score does not correlate with the confidence in the algorithm, even when participants with an ATI score one standard deviation above or below the mean are excluded.

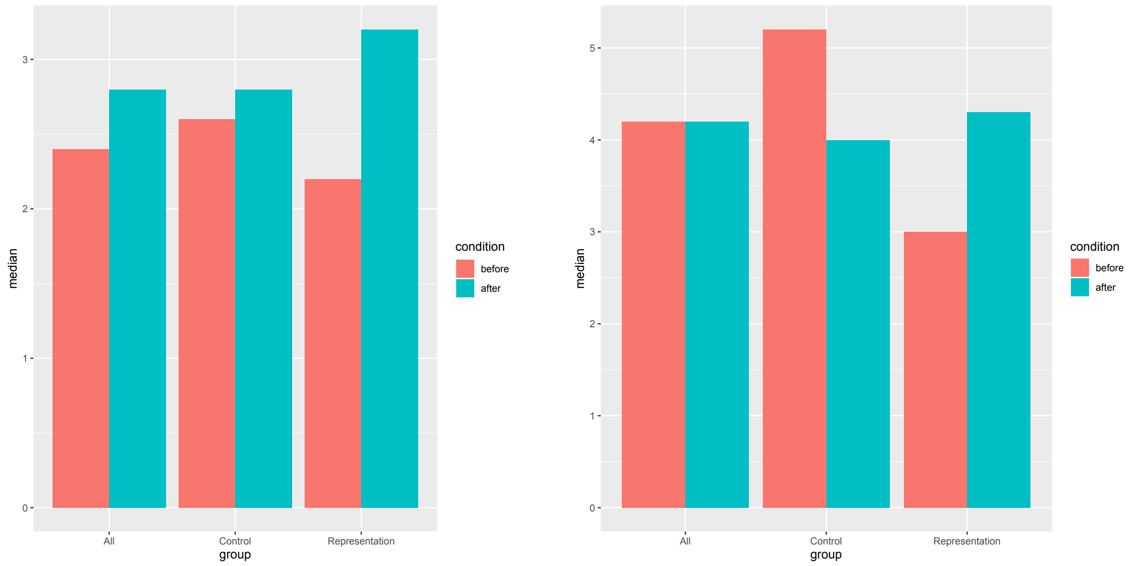
6.2.3 Performance

Figure 9 shows the median absolute deviation from the suggested answers for the group that scored better than the algorithm (left in the figure) during the first half of the experiment and the group that scored worse (right in the figure). The group that scored equal to the algorithm was a lot larger for this experiment in comparison to experiment I (N = 93). We omitted this group from this comparison because the before value is zero and comparing it with the after value would therefore not make sense, as the values can only deviate one way.

The data set of participants who performed better than the algorithm is rather small (N = 46). All participants combined show a significant loss of trust after the break (median before = 2.4, median after = 2.8, Z = -2.43, p = 0.0015 and r = -0.70). This is due to the significant change in the representation group (median before = 2.2, median after = 3.2, Z = -2.42, p = 0.0016 and r = -0.71), because the change in the control group is not significant (median before = 2.6, median after = 2.8, p = 0.18), as expected from seeing figure 9a. However, the effect of the group condition on all three response variables is not significant (all $p > 0.29$). This could be caused by the small sample size that remained in this data set, or the medians between the groups are significantly different but the pairings are not.

The drop in the representation group is in line with the results of experiment I and the expected outcome. The algorithm provides two estimations that are far above the average fault and on top of that the participants score better, which is a valid reason to use their own judgement more in the second half. Therefore, it makes sense that they deviate more from the algorithm. In this experiment the control group also trusts their own capabilities more than the algorithm when they have seen it err. While experiment I showed a decrease in the deviation and thus a rise in trust, the drop in trust of the control group in experiment II is more in line with the algorithmic aversion effect. The difference could be explained by the fact that the algorithm in experiment I performed better during the second half. The decline in deviation could therefore be due to the fact that guessing closer to the algorithm's suggestion provided better answers and hence participants did not purposely deviate less from the algorithm, they just scored better.

Participants who performed worse than the algorithm reacted to this differently on average based on which group they were in. The control group deviated less from the algorithm in the second half and the representation group deviated more. All of these changes were not significant (all $p > 0.17$). The group condition did not have a significant impact on the response variables in the models, although the values before and after were almost significant (before: $p > 0.08$, after $p = 0.06$ and delta: p - 0.71). The fact that the difference in *delta deviation* is not significant can once again be explained by the fact that the medians are not a good representation of the individual pairings in this experiment. For the representation group in this experiment we can state that the disappointment of seeing results that are far above the average error weighs heavier than the notion that it performed better than the participants on average.



(a) Median of absolute deviations for participants who performed better than the algorithm during the first half. ($N = 46$).

(b) Median of absolute deviations for participants who performed worse than the algorithm during the first half. ($N = 133$).

Figure 9: Median of the mean absolute deviations from the algorithm's suggestions per participant before and after the break, for the two groups and all data.

The participants that scored better than the algorithm on average saw a drop in their performances when they deviated more from the algorithm (see table 6 and figure 9), ending up performing significantly worse than it in the second half (Z = -5.67, $p < 0.001$, $r = 0.41$). Comparably, the worse performing participants saw an increase in performance after the break, which is not significant ($p = 0.16$). For the control group this is clear: they conform more to the algorithm, although far from significant ($p = 0.45$). The fact that the representation group scores better on average after deviating more from the algorithm seems counter-intuitive. However, the difference in means shows that there is an average absolute decrease in performance (before = 8.04, after

Group	N	Mean performance before		Mean performance after	
		Median	SD	Median	SD
Algorithm	1	6	NA	6	NA
Better - All	46	5.2	0.93	6.8	1.9
Better - Control	23	5.2	0.62	6.8	2.3
Better - Representation	23	5.2	1.2	7	1.7
Worse - All	133	7.8	4.8	7.6	4.5
Worse - Control	83	8.4	4.8	8	4.9
Worse - Representation	50	6.9	4.6	6.7	3.3

Table 6: The performances of the participants performing better and worse than the algorithm in experiment II, per group (lower scores are better). The mean performances are the means of the absolute deviations from the right answers per participant.

= 9.00), but this does not reflect in the medians because the standard deviation is lower in the *after* score. The difference in performance for the representation group is almost significant (p = 0.07).

6.2.4 Using the algorithm as an anchoring tool

The difference in deviation of the participants depending on the quality of the algorithm's suggestion that was found in experiment I is not replicated in experiment II. There is no significant difference between the deviations from the good suggestions and the bad suggestions (median good = 2.7, median bad = 3, p = 0.50). The difference in performance between the good and bad suggestions is replicated (median good = 4, median bad = 12, Z = -15.5, $p < 0.001$ and r = -0.707).

Although there is no significant difference between the deviations from the good and bad suggestions, participants score better at the former. It could still be the case that the algorithm's suggestion is useful as anchor point to end up at a better answer than the algorithm, as they perform better than the algorithm on the tasks with good suggestions.

6.3 Discussion

The differences between the medians before and after the break are significant for both groups and all data. The differences were all negative, trusting the algorithm less after the break which replicates the algorithmic aversion effect. Even without mentioning that the algorithm was made by a respectable company, the difference between the groups before the break is significant. This provides evidence against one of the theories raised after the first experiment: that this line in the representation briefing could have given the algorithm a certain authority that participants followed.

Although the data seem to indicate a difference between the representation and the control group in terms of *delta deviation*, the effect is not significant. H1 can therefore not be accepted.

Because the algorithm is quite accurate apart from its obvious mistakes, deviating more from the algorithm reduces the performance. This is exactly what happens for participants who score better than the algorithm in the first half, especially the participants who are also in the representation group. Participants who performed worse than the algorithm in the first half improved by conforming more to it.

Similarly to experiment I, no significant results were found for the ATI variable. There is thus no significant effect of the ATI value on the algorithmic aversion effect or on the difference in trust between groups. H2 and H3 can therefore not be accepted.

Participants score better at tasks for which the algorithm provides a good suggestion (one below the promised mean error of 4.32); they even outperform the algorithm in these tasks. There is no significant difference between the deviations from the good and bad suggestions. This is due to the fact that the participants deviate from the algorithm towards the right answers for the tasks with good suggestions. Their significantly better score might indicate that they do use the algorithm's suggestion as anchor point to reach these good guesses.

7 General discussion

We examined the influence of the representation of an algorithm on the participants' trust in it. The representation group got information about the way the algorithm works and the average fault it has, while the control group performed the experiment without this information.

The representation group starts with significantly more confidence in the algorithm in both experiments, but this drops drastically when the participants in this group see the algorithm err. In experiment I this difference between the groups in terms of *delta deviation* (the average deviation from the algorithm after the break minus the average deviation before the break) is significant. Although the data of experiment II show a substantial difference between the groups, this difference is not significant. This is probably due to the fact that the median values do not correspond with the participants' pairs of values before and after the break. This way the median difference over the whole data set is significant, but the individual pairings per participant (the variable we are interested in) are not.

The significant drop in trust for both groups in experiment II is in line with the algorithmic aversion effect. We did not find a significant drop in trust in experiment I, but there are some caveats. We have reason to believe that people get better at the task in experiment I. If they indeed get better, they deviate less from the algorithm, because the algorithm performs quite well on average. There is also an inequality between the performance of the algorithm before and after the break which could influence the deviations from it. These differences could have made the effect significant for experiment I.

We did not find a significant effect of the participants' technical affinity on the results. There is no precedent for this interaction and therefore it is not surprising that we did not find it. It was purely our curiosity that led us to include it in the models. While it is likely that there just is no effect of technical affinity on the trust, some other possibilities have to be discussed. It could be the case that the chosen test for technical affinity is not an accurate representation of it. Another explanation is that everybody uses computers enough in their daily lives to score fairly highly on the test, reducing its variance. The average score was always fairly high (approximately 4/6). The recruitment of participants on Amazon MTurk increases this imbalance by excluding the computer-illiterate population.

Splitting the data based on whether the participants scored better or worse than the algorithm showed that those who performed worse conform more to the algorithm. We found the opposite effect in the better performing group. The scores of the better performing groups drop by deviating more from the suggestions, since the algorithm is quite accurate in most cases. In contrast, the worse performing group gets better results by conforming more to the algorithm.

A comparison of the deviations from good and bad algorithmic suggestions suggests that the participants use the algorithm as an anchoring tool after the break. In experiment I there is a significant difference between the deviations from the good and bad suggestions, indicating that the participants pay attention to the suggestions and use them as they see fit, but keep critical of it. The same difference is not significant in experiment II. This could however be explained by the fact that participants outperform the algorithm on tasks with good suggestions. This would imply that they do deviate the same amount from good and bad suggestions, but in the case of the good suggestions they deviate towards the right answer. This is a good example of hybrid intelligence, where an algorithm augments the performance of humans but does not replace them.

7.1 Limitations and future work

To limit the length of the survey, it contains only ten estimation tasks in total. Long surveys in general lead to worse attention of the participants, and on MTurk in particular it has shown to lead to worse data[32]. The limited number of tasks might make the introduced mistakes too obvious. In a longer experiment there could have been three phases, where the first phase would be used to earn the trust of the participants by providing accurate suggestions. An extra break page and phase could also be added after the current experiment, to see whether the trust changes after another round of tasks. The participants who scored better than the algorithm in the first half of the experiments in this paper performed a lot worse during the second half because they deviated more from the algorithm. A question that remains is whether they will deviate from the algorithm less again after a second break where they see that the algorithm actually performs quite well on average.

The COVID-19 pandemic made offline experiments impossible. The representation factor may be more profound when the participants actually see an algorithm in action instead of digesting a page of information about it. An offline equivalent of the "M&Ms in a vase" experiment would show the representation group the vase of M&Ms with a webcam rotating around it to take the necessary pictures. The control group would just get the results presented on a screen. If one extra information page already causes a significant change in answers, a live demo could be even more influencing.

It would also be interesting to test different briefings for the representation group. In these experiments the algorithm overpromised and underdelivered. People might react totally different when the algorithm performs better than was promised.

Finally, the anchoring effect of the suggestions is very interesting. Future work could vary the algorithm's suggestions to evaluate the differences in the way participants use the suggestions. This requires a clever research design in which the right effect can be measured in the participants' reactions, under a limited number of conditions to reduce the number of participants needed. Measuring the representation and control groups using a few different suggestions by the algorithm leads to an enormous amount of participants needed.

8 Acknowledgement

Firstly, I, Sander, would like to thank Maarten Lamers for supervising this project from start to finish, giving it direction where it was needed and narrowing it down where it was wise. I also would like to thank him for his abundance of comments on my drafts, which improved this paper with each iteration. I want to thank Maarten Bos for his large time investment to share his insight into the research on decision making, his statistics lessons and the opportunity to use the Snap Inc. Amazon MTurk account to recruit participants, all of which made this process a lot easier and a lot more pleasant. Additionally, I would like to thank Berkeley Dietvorst, who provided us with the supplementary resources to replicate his "Airline Passengers" experiment, and Max van Duijn, who helped give this project direction at the start by providing very useful book recommendations about the theoretical background of decision making.

I would like to thank Eliška for her mental support throughout the process as well as her keen eye for photography, which enabled us to run experiment I online in uncertain times. Thanks to Gabe, Gaia and Romy for discussing our projects together and sharing our experiences from the start of our studies until graduation. In addition, thanks to Gabe for proofreading this thesis.

I would like to dedicate this thesis to my parents, Teus and Monique, who are not able to witness the completion of my master's degree. This work would not have been possible without the various bases of knowledge they each possessed and shared with me, the curiosity they instilled in me and the perseverance they showed me at moments when life gets hard.

References

[1] Niiler, E. (2019). Can AI Be a Fair Judge in Court? Estonia Thinks So. Web. 25 September 2019. https://www.wired.com/story/can-ai-be-fair-judge-court-estonia-thinks-so/

[2] Longoni, C., Bonezzi, A. & Morewedge, C. (2019). Resistance to Medical Artificial Intelligence. Journal of Consumer Research. 46. 10.1093/jcr/ucz013.

[3] Promberger, M. & Baron, J. (2006), Do patients trust computers?. J. Behav. Decis. Making, 19: 455-468. doi:10.1002/bdm.542

[4] Li, L., Ota, K. & Dong, M. Humanlike Driving: Empirical Decision-Making System for Autonomous Vehicles, in IEEE Transactions on Vehicular Technology, vol. 67, no. 8, pp. 6814-6823, Aug. 2018, doi: 10.1109/TVT.2018.2822762.

[5] Kugler, L. (2018). AI Judges and Juries: Artificial intelligence is changing the legal industry. Communication of the ACM. Vol. 61, no.12, December 2018. DOI:10.1145/3283222

[6] Fry, H. (2018). Hello World: Being Human in the Age of Algorithms (pp. 57-91). Transworld Publishers.

[7] Dietvorst, B., Simmons, J., & Massey, C. (2015). Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err. Journal of Experimental Psychology: General, 144(1), 114-126.

[8] Dijksterhuis, Ap. (2004). Think Different: The Merits of Unconscious Thought in Preference Development and Decision Making. Journal of personality and social psychology. 87. 586-98. 10.1037/0022-3514.87.5.586.

[9] Dijksterhuis, A, Bos, M., Nordgren, L. & Baaren, R. (2006). On Making the Right Choice: The Deliberation-Without-Attention Effect. Science (New York, N.Y.). 311. 1005-7. 10.1126/science.1121629.

[10] Kahneman, D. (2012). Thinking, fast and slow. London: Penguin.

[11] Austin, W. & Williams, T. A. III. (1977). A Survey of Judges' Responses to Simulated Legal Cases: Research Note on Sentencing Disparity, 68 J. Crim. L. & Criminology 306.

[12] Dhami, M. & Ayton, P.. (2001). Bailing and Jailing the Fast and Frugal Way. Journal of Behavioral Decision Making. 14. 141 - 168. 10.1002/bdm.371.

[13] Furnham, A., & Boo, H. C. (2011). A literature review of the anchoring effect. The journal of socio-economics, 40(1), 35-42.

[14] Wansink, B., Kent, R. & Hoch, SJ. (1998). An Anchoring and Adjustment Model of Purchase Quantity Decisions. Journal of Marketing Research. 35. 71-81. 10.2307/3151931.

[15] Galton, F. Vox Populi . Nature 75, 450–451 (1907). https://doi.org/10.1038/075450a0

[16] Lee, M.K (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. Big Data & Society, pages 1-16

[17] Hertz, N. (2016) Non-human Factors: exploring conformity and compliance with non-human agents. Dissertation at George Mason University.

[18] Castelo, N., Bos, M. W., & Lehmann, D. R. (2019). Task-Dependent Algorithm Aversion. Journal of Marketing Research, 56(5), 809–825. https://doi.org/10.1177/0022243719851788

[19] Logg, J., Minson, J. & Moore, D. (2019). Algorithm appreciation: People prefer algorithmic to human judgment, Organizational Behavior and Human Decision Processes, Volume 151, 2019, Pages 90-103, ISSN 0749-5978, https://doi.org/10.1016/j.obhdp.2018.12.005.

[20] Dietvorst, B., Simmons, J., Massey, C. (2018) Overcoming Algorithm Aversion: People Will Use Imperfect Algorithms If They Can (Even Slightly) Modify Them. Management Science 64(3):1155-1170. https://doi.org/10.1287/ mnsc.2016.2643

[21] Kizilcec, R. (2016). How Much Information?: Effects of Transparency on Trust in an Algorithmic Interface. 2390-2395. 10.1145/2858036.2858402.

[22] Edison, S., Geissler, G. Measuring attitudes towards general technology: Antecedents, hypotheses and scale development. J Target Meas Anal Mark 12, 137–156 (2003). https://doi.org/10.1057/palgrave.jt.5740104.

[23] Cacioppo, J. & Petty, R. (1982). The Need for Cognition. Journal of Personality and Social Psychology. 42. 116-131. 10.1037/0022-3514.42.1.116.

[24] Schmettow, M. & Drees, M. (2014). What drives the geeks? Linking computer enthusiasm to achievement goals. 10.14236/ewic/hci2014.29.

[25] Karrer, K., Glaser, C., Clemens, C. & Bruder, C. (2009). Technikaffinität erfassen – der Fragebogen TA-EG. In A. Lichtenstein, C. Stößel und C. Clemens (Hrsg.), Der Mensch im Mittelpunkt technischer Systeme. 8. Berliner Werkstatt Mensch-Maschine-Systeme (ZMMS Spektrum, Reihe 22, Nr. 29, S. 196-201). Düsseldorf: VDI Verlag GmbH.

[26] Richter, T. & Naumann, J. & Horz, H. (2010). A Revised Version of the Computer Literacy Inventory. Zeitschrift für Pädagogische Psychologie. 24. 23-37. 10.1024/1010-0652/a000002.

[27] Franke, T., Attig, C., & Wessel, D. (2018). A Personal Resource for Technology Interaction: Development and Validation of the Affinity for Technology Interaction (ATI) Scale. International Journal of Human–Computer Interaction, 1–12. doi:10.1080/10447318.2018.1456150

[28] Kruger, J. & Dunning, D. (2000). Unskilled and Unaware of It: How Difficulties in Recognizing One's Own Incompetence Lead to Inflated Self-Assessments. Journal of Personality and Social Psychology. 77. 1121-34. 10.1037//0022-3514.77.6.1121.

[29] Salomons, N. (2018) Humans Conform to Robots: Disambiguating Trust, Truth, and Conformity. In Proceedings of HRI'18, Chicago, IL, USA, March 5th–8, 2018, 9 pages. https://doi.org/10.1145/3171221.3171282

[30] Hertz, N. & Wiese, E. (2018). Influence of Agent Type and Task Ambiguity on Conformity in Social Decision Making. Proceedings of the Human Factors and Ergonomics Society 2016 Annual Meeting. DOI 10.1177/1541931213601071

[31] Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk. Perspectives on Psychological Science, 6(1), 3–5. doi:10.1177/1745691610393980

[32] Buhrmester, M. D., Talaifar, S., & Gosling, S. D. (2018). An Evaluation of Amazon's Mechanical Turk, Its Rapid Rise, and Its Effective Use. Perspectives on Psychological Science, 13(2), 149–154. doi:10.1177/1745691617706516

[33] Rouse, S. V. (2015). A reliability analysis of Mechanical Turk data. Computers in Human Behavior, 43, 304–307. doi:10.1016/j.chb.2014.11.004

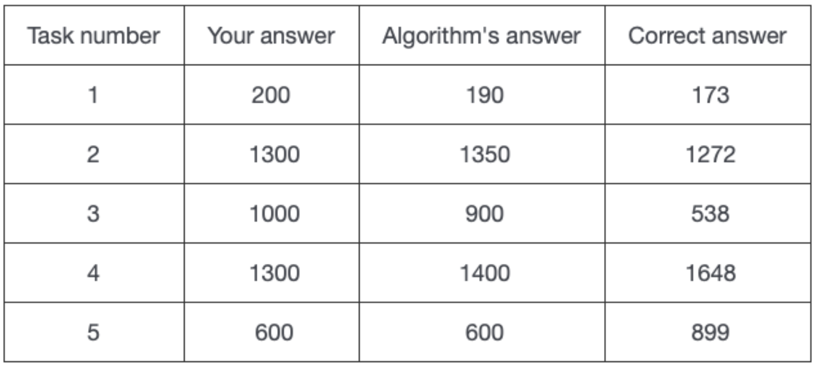
[34] Paolacci, G., & Chandler, J. (2014). Inside the Turk. Current Directions in Psychological Science, 23(3), 184–188. doi:10.1177/0963721414531598

Appendix A: The Affinity for Technology Interaction scale

The questions that are part of the ATI test. The questions are answered using a 6 point Likert scale. The negatively worded questions (3,6 and 8) are reversed before the final score is calculated by taking the mean of the question scores, resulting in a score between 1 and 6 (where 1 is low and 6 is high).

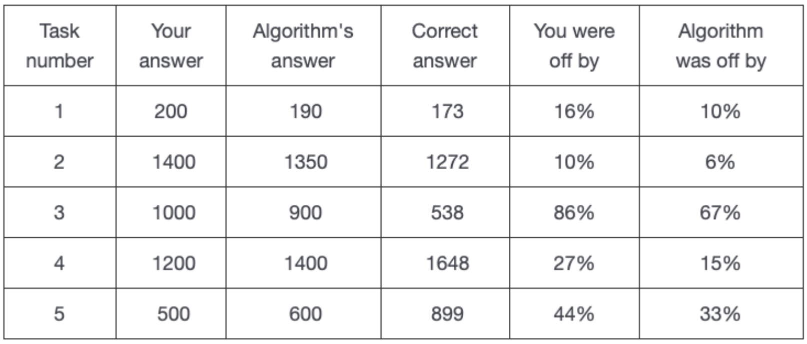
1. I like to occupy myself in greater detail with technical systems.
2. I like testing the functions of new technical systems.
3. I predominantly deal with technical systems because I have to.
4. When I have a new technical system in front of me, I try it out intensively.
5. I enjoy spending time becoming acquainted with a new technical system.
6. It is enough for me that a technical system works; I don't care how or why.
7. I try to understand how a technical system exactly works.
8. It is enough for me to know the basic functions of a technical system.
9. I try to make full use of the capabilities of a technical system.

Appendix B: The two tables used on the break page



Task number	Your answer	Algorithm's answer	Correct answer
1	200	190	173
2	1300	1350	1272
3	1000	900	538
4	1300	1400	1648
5	600	600	899

Figure 10: The break page used in experiment I without extra columns



Task number	Your answer	Algorithm's answer	Correct answer	You were off by	Algorithm was off by
1	200	190	173	16%	10%
2	1400	1350	1272	10%	6%
3	1000	900	538	86%	67%
4	1200	1400	1648	27%	15%
5	500	600	899	44%	33%

Figure 11: The break page used in experiment I with extra columns.