Influence of Algorithms: Empirical study on the influence of algorithms' reliability and transparency on the users' decision-making process

Sören M. Schröder (399838)

Computer Science

Individual and Technology, RWTH Aachen University

Empirical Investigation of Communication in Human-Robot-Interaction

Prof. Dr. Astrid Rosenthal-von der Pütten

28. August 2020

Abstract

Algorithms are a part of our everyday life. So-called algorithm decision systems (ADS) are systems that are designed to make a decision, bases on several input data. They are used to support professionals in the decisions they have to make. For developing these systems often the technology of machine learning is used, which uses huge datasets to train an algorithm to detect certain patterns. One of the most remarkable achievements was the development of an algorithm that can detect melanoma in an early stage with similar success rates than medical professionals. When using such a system medical professionals might be influenced in their decision by the ADS. To investigate how transparency and reliability of the ADS influence the professionals' decision-making process we performed an empirical online study. In the study, medical students should assess pictures of spots with the help of a mocked-up ADS. During the experiment factors like fairness, confidence, and conformity were measured. We found that people do not confirm with unreliable ADS and that the gain in understanding of the algorithm by using it is influenced by its transparency. Furthermore, our findings indicate that people's self-reported confidence in the algorithm diverges from how they behave.

Keywords: algorithmic decision making, algorithms, transparency, reliability, fairness, conformity, confidence, understanding, skin cancer, melanoma

Table of Contents

1. Introduction

During the last decade, algorithms gained a more important and omnipresent role in our society. In general, an algorithm is a series of calculations that receives a series of inputs and transform it into a series of outputs (Cormen, Leiserson, Rivest, & Stein, 2013). For us, the algorithms are interesting which are defined by Cormen et al. (2013) as correct algorithms. They define them as algorithms that stop for a certain input and delivers a correct result and therefore solve the calculation. Another characteristic of an algorithm is that for the same input always the same corresponding output is returned (Rogers, 1967), in other words, the algorithm is deterministic.

1.1 Algorithmic Decision Making

Algorithms can not only be manually written by humans and then get executed. They can also be generated by analyzing data and recognizing patterns in it. This approach (often referred to as *machine learning*) can be used to develop algorithm decision systems (ADS), which are involved in the process of decision making (Castelluccia & Le Métayer, 2019). They describe that the involvement of humans can be stated in a spectrum that ranges from systems that provide advice for a human who is finally responsible for the decision to systems that make decisions full automatically.

How algorithms on the autonomous end of the described spectrum can be used for job applications was shown by Wang, Harper, & Zhu (2020) in their work about how the fairness of algorithmic decision systems is perceived by concerned individuals. In their work, they mocked a system that behaved as if it could promote workers on a crowdsourcing workplace (an online platform where workers get paid for fulfilling micro tasks (Wang et al., 2020)), based on worker's data provided to the algorithm. Even if this study did not use a real algorithm, because their research focused on the perceived fairness, that ADS for job applications will eventually become reality. In other fields, ADS are already in use. One of these is the COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) software developed by Northpoint, Inc. (Kirkpatrick, 2017). The tool is used to decide whether a defendant is allowed to be free on bail or should be kept in (Hao & Stray, 2019). Also in the field of medicine algorithmic decision systems already gathered attention. Esteva et al. (2017) developed an ADS that was able to detect skin cancer by analyzing images of the corresponding skin areas. Thereby they

achieved success rates similar to them of experts they have tested. This topic is of special interest since skin cancer is the most common malignant disease among humans (Esteva et al., 2017). A dermatologist diagnoses melanomas primarily visually by applying the ABCDE method (Esteva et al., 2017). ABCDE is an abbreviation for different characteristics of a spot which can be used as an aid to detect melanoma in an early stage (Rigel, Friedman, Kopf, & Polsky, 2005). The characteristics of the spot which are used by the method to assess them are asymmetry, border irregularity, color, diameter, and evolution (of the spot over time). The characteristics of the spot which are used by the method to assess them are **a**symmetry, **b**order irregularity, **c**olor, **d**iameter, and **e**volution (of the spot over time).

The development of algorithms as presented by Esteva et al. is most likely to be continued. Either to provide easy access to early skin cancer detection for many humans by using smartphones (Esteva et al., 2017) or to support dermatologists in their work. For the latter, it would be interesting to investigate several effects that emerge from their use. Therefore, this study investigates the use of ADS in the context of medicine on the example of the task of assessing spots on the skin to being melanoma. We set a focus on the transparency and reliability of the ADS and how they influence the decision-making process of medical professionals. In the following, we use the algorithm as a synonym for ADS, since we did not introduce the term ADS during the study because it was not necessary to know for the participants.

1.2 Transparency

In the application area of jurisdiction, the final decision is made by a judge (Kirkpatrick, 2017) and the same would apply when using ADS in a medical context. For COMPAS we know, that it is biased against particular subgroups (Angwin, Larson, Mattu, & Kirchner, 2016) and similar problems will likely occur in other systems for different application areas. Since the user of an ADS (e.g. judges or doctors) usually do not have the necessary technical knowledge to understand how the algorithm comes to its decision, the level of transparency the algorithm provides could be from significant interest regarding the question of how they are going to be used. Kizilcec (2016) showed that the provided transparency level of the ADS influences the trust in it. So we state our *Research Question 1: How is the user's decision-making influenced by the transparency of the involved algorithm?*

1.3 Fairness

Prior research could not found a relation between transparency and perceived fairness in the context of ADS (Wang et al., 2020). Two things might have led to this result. First, the operationalization of transparency by stating that the algorithm was developed transparent or not might have been a too weak stimulus to measure an effect. Second, their participants were directly affected by the result which might interfere with the perceived fairness. By overcoming these two problems by giving a stronger stimulus and using a setting where the user is not directly affected we state *Hypothesis 1: Transparency of the algorithm is positively associated with the perception of fairness*.

1.4 Reliability

Another factor that might influence the perception of the algorithm is how good it performs (i.e. how reliable it is). Studies have shown that the performance of an algorithm can influence how it is perceived by the user. Thereby, the performance was either perceived directly by seeing that the algorithm fails (Dietvorst, Simmons, & Massey, 2015) or by having the information that the algorithm is known to make errors by providing information about error rates and biases (Wang et al., 2020). Since it is difficult to prevent errors during the development of software in general and for ADS in particular, because an already biased dataset can lead to a biased algorithm (Hao & Stray, 2019), it is important to know how the users of an ADS are going to cope with the shortcomings of the algorithm and how it influences their decision-making process. So this study also investigates *Research Question 2: How is the user's decision-making influenced by the reliability of the algorithm involved?*

1.5 Confidence

When people recognize that an algorithm makes an error, they will lose confidence in the algorithm (Dietvorst et al., 2015). In their study, they found out that the loss of confidence is even higher as if the same mistake was made by a human. We suggest that this will also be the case if the participant not only will perceive a bad performance, which led us to *Hypothesis 2: The users will show less confidence in decisions made by an unreliable algorithm*, but also if they know that the algorithm will make errors due to a known error rate. Wang et al. (2020) reported that a known bias influenced the user perception of the algorithm. So we state *Hypothe-*

sis 3: The confidence in the algorithm would be less when the algorithm's error rates are known as compared to when no information about error rate is provided.

1.6 Conformity

Since the confidence in the algorithm is expected to decrease for unreliable algorithms we also expect the users to deviate with their predictions form the ones stated by the algorithm. This lack of conformity gave us *Hypothesis 4: The user's predictions (of the probability of the nevus being melanoma) will deviate more from the predictions of a less reliable algorithm*

1.7 Understanding

When using ADS as aid it would be important that the users understand how the algorithm used and how the results need to be interpreted. This is necessary to use the algorithm's output for their own decisions. Transparency could be a key factor for understanding and ADS. Wortham, Theodorou, & Bryson (2017) showed that providing an insight into the decision-making process of an algorithm can increase the user's understanding. Since insight into an algorithm is a way of making the algorithm more transparent we expect that this will also be the case when providing transparency not during the use but in advance in form of information about how the algorithm was developed. Therefore, we state our *Hypothesis 5: The understanding of the algorithm, before it is used, would be higher for the algorithm with high transparency as compared to one with low transparency*. Besides, we also want to have a look into how the understanding of the algorithm is changed by working with it (i.e. using the algorithm). This led us to *Research Question 3: How does the use of algorithms in the decision-making process change its understanding?*

2. Method

To answer the stated hypotheses and research questions an empirical study was performed as an online experiment, which had the advantage that it was easier to distribute it to more potential participants than conducting the experiment in the lab. The survey was developed with SoSciSurvey[1] and can be found in Appendix A (see page 22).

1. https://soscisurvey.de

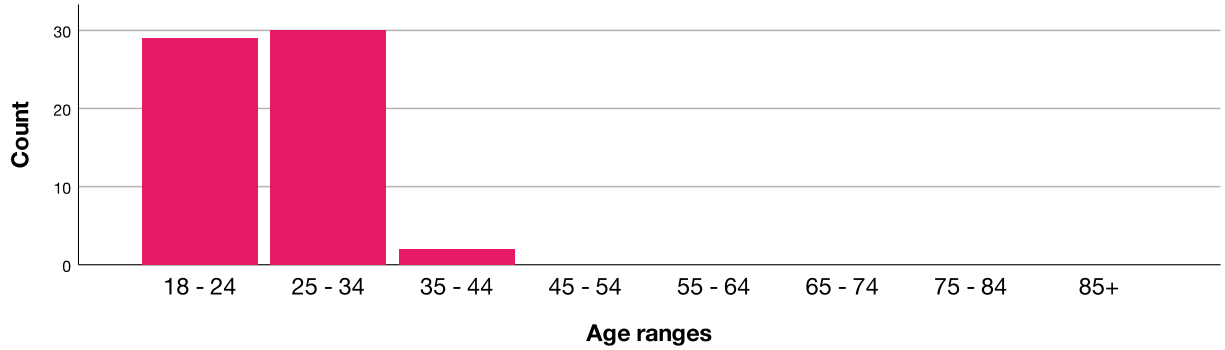
2.1 Participants

In total 61 people participated in the study. Since there was no interest to capture the exact age the participants were asked to select an age group. To enable the participants to give informed consent, it was necessary to limit the age of the participants to at least 18 years. The results show that the participants were not older than 44 years and most participants are 25 to 34 years old (see Figure 1). The absence of older participants was expected and was due to the recruitment strategy. The taken convenience sample was recruited in several ways. The survey was posted in Facebook groups of medical students, send to faculties of medical students with the demand to distribute the survey, and it was posted on online bulletin boards of medical faculties. Furthermore, personal contacts were asked to participate and distribute the survey, and a class coordinator from the medical faculty of the RWTH Aachen University distributed the survey to their students.

The participants were asked to only participate if they know about of applying the ABCDE-Method, which is the reason the recruitment was limited to the described ways. This requirement was stated in the introduction and verified during the survey by asking where the participants have learned about the method and what their educational background is. From this, we could see that 52 of our participants already knew the ABCDE method. Nevertheless, we have not removed the other participants from the sample since the already low number of participants and the then resulting imbalance of the conditions. As compensation for their time, the participants had the option to register their mail address, which was stored separated from the experimental data, to a lottery with which they had the chance to win one of five Amazon gift cards with a total value of 150 Euro (1 x 50 Euro und 4 x 25 Euro).

Figure 1

Participants reported age



Note. Most participants are between 18 and 34 were they quite evenly distributed below 25 and above 24. Only a few participants were between 35 and 44. No one was 45 or above.

2.2 Procedure

The study was designed as a 2x2 between-subjects online experiment, where the participants were assigned randomly to one of each condition for both factors. The two factors, our independent variables, were the *transparency* of the algorithm (low transparency, high transparency) and the *reliability* of the algorithm (unreliable, reliable). This design gave us the ability to investigate on the one hand side the effect of the transparency of the algorithm on the perceived fairness (Hypothesis 1), confidence (Hypothesis 3), and the understanding of the algorithm before using it (Hypothesis 5), and on the other hand the effects of the algorithm's reliability on the confidence (Hypothesis 2) and the deviation of the user's prediction from the algorithm (Hypothesis 4). The study consists of three parts. Starting with an introduction part containing information about the experiment, basic knowledge, and a questionnaire, followed by the task and its explanation. and completed by several questions after the task, regarding the perception of the algorithm, demographics, and general questions regarding the use of algorithms.

On the introduction page, visitors were informed about the conditions under which they can participate in the study (participation was anonymous, needed time was about 25 minutes, the rough structure of the experiment, a brief description of the task, etc.). They were also informed that the purpose of the study is to evaluate an algorithm for detecting melanoma, which was developed at the computer science department at RWTH. This deception was resolved at

the end of the experiment, but necessary to keep the participants uninformed during the experiment to prevent bias by knowing the actual research goals. Further, they were informed about the optional lottery. Afterward, their attitude towards algorithms, in general, was measured by asking 28 rotated questions from which 14 were asked as reversed items on a 7-point Likert scale from "completely disagree" to "completely agree" (e.g. "Algorithms should not make morally difficult decisions." and "Algorithms apply the same scale to everyone."). The participants were asked to read basic information about the ABCDE-Method to provide them a short recap and ensure some common basic knowledge level. To ensure this, a knowledge check was performed before continuing.

The second part started with the explanation of the task and an example of the later provided cases, which consists of a picture showing a section of human skin with a nevus, information about the symptoms, and an assessment by the algorithm (see Figure 2). The algorithm states to which degree (0% – 100%) it assesses the nevus as a melanoma. Depending on the transparency condition the participants received very little information about how the algorithm was developed (low transparency) or a detailed explanation of how the algorithm works, how it was trained, and how its performance was tested. To ensure the information was red attentive the participants had to answer one (low transparency) or three (high transparency) corresponding questions correctly. We accepted the different amount of questions since we wanted to ensure that in the high transparency condition the whole information was red and at the same time we wanted to keep the information for the low transparency condition as little as possible. We do not expect this to be a significant confounding factor in our study. Afterward, the participants evaluated this explanation and stated how confident they feel in using the algorithm.

Figure 2

Example of a case showed during the experiment

Bitte sehen Sie sich das Bild an und lesen Sie die untenstehenden Informationen sorgfältig durch. Und beantworten Sie die folgenden Fragen:

- **Der Algorithmus sagt voraus, dass es sich bei diesem Muttermal mit einer Wahrscheinlichkeit von 12% um ein Melanom handelt.**
- **Der Fleck auf der Haut ist im Laufe der Zeit verblasst.**

Note. The image shows a *correct negative* case. The shown spot is no melanoma and the assessment of the algorithm reflects that.

For each of the 15 provided cases the participants were asked to state there own prediction of how likely the nevus is a melanoma (0% – 100%) and if they would perform a biopsy (*yes, no)*. We also asked to state on a Likert scale (1 = *Not at all sure*, 5 = *Very sure*) how sure

they are with their own decision (operationalization of self-confidence) and also on a Likert scale (1 = *Not at all reliable, 5* = *Very reliable*) how reliable they would rate the assessment of the algorithm (operationalization of confidence in the algorithm). The 15 cases were divided into 5 negative cases (clearly no melanoma), 5 positive cases (clearly melanoma), and 5 ambiguous (not clearly) cases. The classification and images were provided by a dermatologist from the RWTH Aachen University. For the unreliable condition, one of the positive and two of the negative cases were changed to a false positive (clearly no melanoma, but algorithm state the opposite) and 2 false negatives (vise versa). The reliable condition only got the 15 not wrong assessed cases. We decided not to add any more mistakes to avoid destroying the confidence in the algorithm, which might have led to ignoring the algorithm at all. Also then the error rate would have diverged too much from the rate stated in the algorithm description what could make the participants suspicious. Since a false negative assessment has severe implications for the patient, we decided to add one more instead of an additional false positive.

After performing the task the participants answered a final questionnaire which asked again how confident they feel to use the algorithm. We also asked them to evaluate the algorithm (e.g. "I have largely ignored the algorithm in my decisions"), how fair they perceived the algorithm, give a self-estimation how much they were influenced by the algorithm, and they were asked to state if responsible people in the medical system should be supported by such algorithms. They should also sort several areas of use for ADS (e.g. "recommendations in dating apps", "diagnosis of skin cancer") according to their severity and the likelihood that they would agree to use them. Finally, several demographic data were collected (gender, age, fields of studies, setting they learned ABCDE method, etc.). In the debriefing, the occlusion of the *algorithms* nature and the experiment's conditions and which were used in their case was revealed. Furthermore, the participants were asked to visit a dermatologist if they observe conspicuous nevus on their skin.

3. Results

The results of our measurements show that overall there is less impact of transparency and reliability than we expected beforehand. Nevertheless, the collected data gave some insights into how the user's perception of the algorithm is influenced and the resulting decisions and actions.

3.1 Fairness of the algorithm

After performing all 15 cases the participants were asked to rate the fairness of the algorithm on a Likert scale (1 = *very unfair* to 5 = *strongly fair*). The measured difference between the high transparency condition (M = 3.52, SD = 0.72) and the low transparency condition (M = 3.52, SD = 0.72) was quite small and therefore even higher for the low transparency condition. Besides, a two-way ANOVA did not reveal any support for a significant effect of transparency on the perceived fairness of the algorithm, $F(1, 57) = 1.04$, $p = .313$, $\eta p2 = .018$, which indicates that the perceived fairness is not influenced by the algorithm's transparency. So, there is no support for Hypothesis 1.

3.2 Confidence in the algorithm

The confidence in the algorithm was stated by the participants in each case. Among all cases was between *moderate* and *reliable* (M = 3.48, SD = 0.42). A two-way ANOVA did not show any main effect on the confidence in the algorithm. Neither by transparency, nor by reliability, nor by an interaction of both (see table 1). Therefore, no support for Hypothesis 3 and Hypothesis 2 could be found. Nevertheless, the result (p = .140) suggests at least some effect of reliability exists. So, we looked only at the cases were the unreliable algorithm did obvious mistakes. A two-way ANOVA showed that reliability had a main effect on the confidence in the algorithm, even with a medium effect size (see table 1). But the confidence for the unreliable algorithm was close to *moderate* (M = 2.92, SD = 0.72) and only a bit below the confidence into the reliable algorithm (M = 3.35, SD = 0.55).

Table 1

Statistical analysis of confidence in the algorithm

	All cases				Unreliable cases			
Variable	**df**	***F*(1, 57)**	**Significance**	**Partial Eta Square**	**df**	***F*(1, 57)**	**Significance**	**Partial Eta Square**
Transparency	1	0.839	.363	.015	1	2.490	.120	.042
Reliability	1	2.234	.140	.038	1	6.379	.014	.101
Transparency* Reliability	1	0.063	.802	.001	1	0.129	.720	.002

3.3 Prediction Deviation

We used the participants' assessments of the cases to calculated the deviation to the algorithm by taking the absolute difference in percentage points between the assessments of the participants and the algorithm. The deviation can be described as moderate (M = 17.2%, SD = 7.3). For the reliable algorithm, we can report even lower deviations in the assessments (M = 14.25%, SD = 7.36). In contrast the deviation for the unreliable algorithm shows a 5.96 percentage points higher deviation (M = 20.1%, SD = 6.2). A two-way analysis of variance (ANOVA) revealed a large significant effect of reliability on the deviation of assessments, $F(1, 57) = 11.17$, $p = .001$, $\eta_p^2 = .164$. But no significant effect on the deviation was caused by transparency or interaction between transparency and reliability (see table 2). These results support Hypothesis 4.

Table 2

Statistical analysis of prediction deviation

Variable	df	$F(1, 57)$	Significance	Partial Eta Square
Transparency	1	0.41	.522	.007
Reliability	1	11.17	.001	.164
Transparency*Reliability	1	1.31	.267	.022

3.4 Understanding of the algorithm

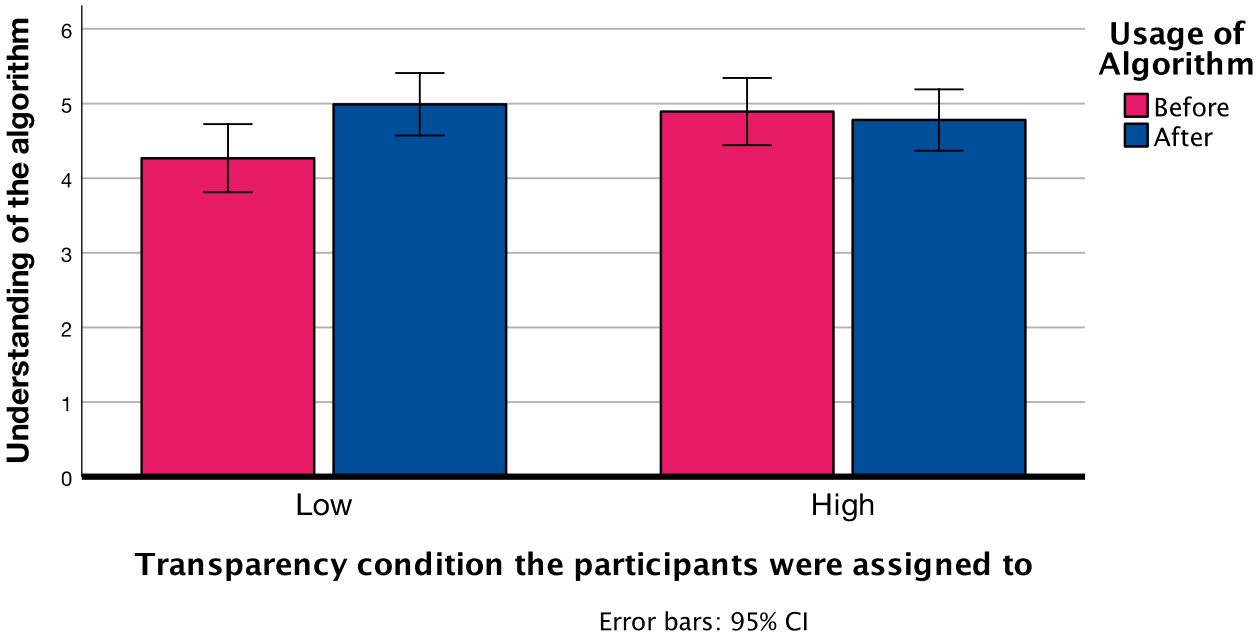
With three items on a Likert scale (1 = *completely disagree* to 5 = *completely agree*) the participants were asked how good they understand *how to use the algorithm*, *how the algorithm works*, and *how good they can grasp the algorithm*. The scale was used after the explanation of the algorithm, right before the cases, and directly after the cases. At both points it provided a high internal consistency (α = .86 and α = .87). Regarding the measurement before assessing the cases, the understanding of the algorithm in the low transparency condition (M = 4.27, SD = 1.41) was below the understanding in high transparency condition (M = 4.89, SD = 1.07). But a two-way ANOVA revealed no significant main effect of transparency on the understanding, $F(1, 57) = 3.81$, $p = .056$, $\eta_p^2 = .063$, which gave no support for Hypothesis 5.

Although, the result indicated that there might be an effect of the transparency on the understanding of the effect. Comparing the change of reported understanding between the two

measurements, before and after the cases, show that for the low transparency group the understanding was higher after using the algorithm's assessment compared to before, with only the explanation. In contrast to the high transparency group, the value dropped a little bit from before to after (see Figure 3). A two-way repeated-measures ANOVA revealed there is a significant effect of the interaction between the transparency and using the algorithm on the understanding of it, $F(1, 57) = 8.32$, $p = .006$, $\eta_p^2 = .127$.

Figure 3

Understanding of the algorithm before and after using it



Note. The understanding of the algorithm has increased for the low transparency group from before to after using the algorithms results. For the high transparency group, the reported understanding decreased a bit from before to after, but not significantly.

4. Discussion

The goal of this study was to get a better insight into how ADS are used by people in general and by medical professionals in particular. We looked at transparency and reliability as factors that influence the perception of the algorithm and thereby how it is used. Therefore, we observed several dependent factors: fairness, confidence, conformity, and understanding.

4.1 Transparency

The study did not found support for the claim that transparency affects the perceived fairness (Hypothesis 1). This result is unexpected since there we know that other research that observed such an effect (Wang et al., 2020). One difference of this study compared to ours is that in our scenario the user of the algorithm was not directly affected by the outcome of the result which might have led to another perception of fairness in general (Greenberg, 1983). Another reason that could explain the differing results could be the kind of decision that was made. The decision made in the study by Wang et al. (2020) was not either right or wrong. To promote someone or not is always a balancing of arguments. In contrast in our study, the assessment of a nevus being a melanoma could be either correct or incorrect. So the perception of fairness in the context of medical assessments might be not connected to the error rate itself. To investigate this topic further we suggest to investigating how different error rates for certain subgroups (e.g. skin color) would influence the perceived fairness. Furthermore, the perception of fairness could be compared between after stating the error rates but before using the algorithm and after using the algorithm to see how the usage of the algorithm changes the perception of fairness.

The confidence of the participants in the algorithm was not significantly affected by the level of transparency, which let us not accept Hypothesis 3. This could be a hint that users, even if they know that the algorithm will make mistakes, are not including this information in their assessment of their confidence into the algorithm. Since we only observed no significant differences in the confidence we have to assume that this kind of transparency is not an influencing factor for users of ADS. On the other hand, there could have been other reasons why we did not observe any influence here. One might be that manipulation by stating error rates was to abstract and therefore too weak to influence the judgment. Since during this time some cognitive load was put on the participants, they had to assess the cases, another reason for observing no effect could be that they forgot about the information that were presented at the beginning of the task section. Further research should investigate if a continuous reminder (e.g. displaying error rates during usage) would affect the confidence in the algorithm.

Regarding the initial assessment of how good the participants understood the algorithm before they used it we could not found support for Hypothesis 5. It seems that a detailed ex-

planation of how the ADS was trained did not lead to significantly better understandings. Since the result was quite close to being significant (p = .056) we assume that our manipulation was not ideal to improve the understanding of the algorithm, but that in general, more information about the algorithm will help users to understand how to use it. To gain a better effect an approach as choose by Wortham et al. (2017) might be more effective. They provided for each decision of the algorithm an insight into how this decision was achieved. Regarding the field of medicine, it might be helpful for medical professionals to have more information the decision the ADS states. It could show how this decision was made and which factors had which amount of influence into it. For further research, we suggest to performing experiments were the participants are provided with more information during the cases to see if this would help them gain a better understanding of the algorithm. To answer Research Question 1, we can detain that transparency seems to have not as much effect on the user's decision-making process as we expected beforehand. On the other hand, we saw that transparency interferes with some kind of learning effect, which suggests that transparency is still important when developing ADS for people who are nonexperts regarding algorithms.

Regarding Research Question 3 we found that there is an interference effect of using the algorithm and transparency. We saw that the understanding increased significantly more for the low transparency condition by using the algorithm compared to the high transparency condition. After using it the knowledge for both groups was on a quite similar level. Even if the understanding of the algorithm before using it was not significantly influenced by the transparency we can see that more information, in the beginning, reduces the increase of understanding during the usage. If the understanding of the algorithm increases during the usage this could mean that especially the first cases a medical professional assess by using ADS might suffer due to the lack of understanding. So we suggest providing information about how the algorithm was trained to help the professionals to early gain an understanding of how to use the algorithm. So we can answer Research Question 3 by stating that the use of the algorithm influences the understanding of it but that also other factors (here transparency but there might be others) have an influence on how this change in understanding expresses.

4.2 Reliability

Besides the influence of transparency, the study investigated reliability as another factor that might influence the usage of ADS. Here we could not find support for Hypothesis 2, which was quite unexpected, since prior research suggested that unreliable ADS will result in less confidence (Dietvorst et al., 2015). A possible explanation could be that the participants did not really perceive the algorithm as performing bad, since they did not get feedback after assessing the cases. For further research, we suggest to change the experiment, such that the assessment of the algorithms quality not relies on the knowledge of the participants. With a more careful selected sample to ensure a higher common basic knowledge this problem could be also solved. Besides this, it is interesting that the participants using the unreliable algorithm deviated significantly from those with a reliable algorithm and still stated the reliability of the algorithm close to *moderate*. This shows a deviation from what they report to how they behave and suggest that self-reporting might not be ideal for measuring confidence in the algorithm (and maybe into oneself, too). Since the behavior is closer to what prior research suggested (Dietvorst et al., 2015) we suggest to measuring confidence in a behavioral manner in future research.

As already mentioned, the reliability of the algorithm influences the conformity of the assessments. These results give us support for Hypothesis 5. From that, we can tell that the participants not just followed the assessments of the ADS but used the ABCDE method to give their assessment. Another reason for the divergence of conformity and confidence could be that the participants just ignored the algorithm. In this study, the stimulus material was on a basic visually level. The image and the results of the algorithm were obviously part of the user interface from SoSciSurvey, which did not make them stand out enough and why they might have been overlooked by the participants. To overcome this weakness and to even provide a more realistic scenario we suggest to mock a user interface of some application and include it into the cover story. This will increase the credibility of the cases and the algorithm.

4.3 Limitations & Further Research

The most obvious and severe limitation of this work is the number of participants. This limitation might be a reason why several measurements showed some differences but often not significant. We think that the lack of significant findings can be solved by performing the study

with more participants. Using a special scenario like detecting skin cancer has on the one hand side the advantage, that it is easier to give the participants a valid reason why they are needed in the study. On the other hand, it reduces the number of people who can participate drastically. We, therefore, propose to study the field of ADS with more general topics. To convince the people that their participation is of importance might be more challenging but it outweighs the advantage of a bigger sample.

By following this suggestion the second limitation of our work, which are the participants without prior knowledge of the ABCDE method, won't have occurred. For studies were a certain knowledge by the participants is important and a small sample sizes are expected, we suggest the following to prevent imbalanced conditions. We saw that asking where the needed knowledge was acquired can be used to filter out not fitting participants. Here it is crucial to check the gathered data on a very regular basis (several times a day) to be able to react to irregular participation and adjust the randomization correspondingly to avoid imbalanced conditions.

In this work, only two levels of reliability were used, which was due to the expected low number of participants. Since the reliability only had a significant influence on the conformation of the assessments, more levels of *bad* reliability would help to gain more insight. Several levels with an even higher number of wrong assessed cases would show how these influence e.g. the confidence in the algorithm. Also, the strength of the error could have been varied. Future research could investigate these more differentiated levels of error to gain knowledge of how different types of errors influence the use of ADS.

5. Conclusion

We showed in this study how the perceptions of medical professionals of ADS is influenced by its transparency and reliability. By manipulating these and measuring other factors (fairness, confidence, conformity, and understanding) we could gain some insight into the user's perception. Besides several not significant influences of the two factors, the main findings in the study were that people do not agree with unreliable algorithms when they made mistakes and that the understanding of these algorithms while using them is influenced by the

amount of information about the algorithm are provided. We also found an interesting divergence between how confident people are in an algorithm and how they behave while using it.

6. Acknowledgments

I would like to thank Sourabh Zanwar for the great cooperation during the semester in this project. Furthermore, I want to thank Prof. Dr. Astrid Rosenthal-von der Pütten and Nikolai Bock for their supervision of this project and provide their help to us throughout the seminar.

7. References

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias. *ProPublica, May*, *23*, 2016.

Castelluccia, C., & Le Métayer, D. (2019). *Understanding algorithmic decision-making: Opportunities and challenges*. European Parliament. https://www.doi.org/10.2861/536131

Cormen, T. H., Leiserson, C. E., Rivest, R., & Stein, C. (2013). *Algorithmen - Eine Einführung*. Walter de Gruyter GmbH & Co KG. https://doi.org/10.1515/9783110522013

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, *144*(1), 13. https://repository.upenn.edu/cgi/viewcontent.cgi?article=1392&context=fnce_papers

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, *542*(7639), 115-118. https://doi.org/10.1038/nature21056

Greenberg, J. (1983). Overcoming Egocentric Bias in Perceived Fairness Through Self-Awareness. *Social Psychology Quarterly*, *46*(2), 152-156. https://doi.org/10.2307/3033852

Hao, K., & Stray, J. (2019). *Can you make AI fairer than a judge? Play our courtroom algorithm game*. Retrieved 2020-08-16 from https://www.technologyreview.com/2019/10/17/75285/ai-fairer-than-judge-criminal-risk-assessment-algorithm/

Kirkpatrick, K. (2017). It's not the algorithm, it's the data. *Commun. ACM*, *60*(2), 21-23. https://doi.org/10.1145/3022181

Kizilcec, R. F. (2016). How Much Information? Effects of Transparency on Trust in an Algorithmic Interface. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2390-2395. https://doi.org/10.1145/2858036.2858402

Rigel, D. S., Friedman, R. J., Kopf, A. W., & Polsky, D. (2005). ABCDE—An Evolving Concept in the Early Detection of Melanoma. *Archives of Dermatology*

Arch Dermatol, *141*(8), 1032-1034. https://doi.org/10.1001/archderm.141.8.1032

Rogers, H. (1967). *Theory of Recursive Functions and Effective Computability*. https://doi.org/10.1137/1011079

Wang, R., Harper, F. M., & Zhu, H. (2020). Factors Influencing Perceived Fairness in Algorithmic Decision-Making: Algorithm Outcomes, Development Procedures, and Individual Differences. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1-14. https://doi.org/10.1145/3313831.3376813

Wortham, R. H., Theodorou, A., & Bryson, J. J. (2017). *Robot Transparency: Improving Understanding of Intelligent Behaviour for Designers and Users Towards Autonomous Robotic Systems*. In Y. Gao, S. Fallah, Y. Jin, & C. Lekakou(pp. 274-289). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-64107-2_22

8. Appendix A

Will be appended in final Version.