June 2021

Can AI reduce motivated reasoning in news consumption? Investigating the role of attitudes towards AI and prior-opinion in shaping trust perceptions of news

Magdalena WISCHNEWSKI [a] and Nicole KRÄMER [a]
[a] *University of Duisburg-Essen*

Abstract. A central role in understanding the interaction between humans and AI plays the notion of trust. Especially research from social and cognitive psychology has shown, however, that individuals' perceptions of trust can be biased. In this empirical investigation, we focus on the single and combined effects of attitudes towards AI and motivated reasoning in shaping such biased trust perceptions in the context of news consumption. In doing so, we rely on insights from works on the machine heuristic and motivated reasoning. In a 2 (author) x 2 (congruency) between-subjects online experiment, we asked N = 477 participants to read a news article purportedly written either by AI or a human author. We manipulated whether the article represented pro or contra arguments of a polarizing topic, to elicit motivated reasoning. We also assessed participants' attitudes towards AI in terms of competence and objectivity. Through multiple linear regressions, we found that (a) increased perceptions of AI as objective and ideologically unbiased increased trust perceptions, whereas (b), in cases where participants were swayed by their prior opinion to trust content more when they agreed with the content, the AI author reduced such biased perceptions. Our results indicate that it is crucial to account for attitudes towards AI and motivated reasoning to accurately represent trust perceptions.

Keywords. trust in AI, AI journalism, machine heuristic, motivated reasoning

1. Introduction

With the rapid development of AI technology, its non-determinism and complexity, as well as AI permeating ever more aspects of everyday life, questions about its acceptance and adaption arise. To that end, special importance has been devoted to the role of trust in AI. Originating in earlier works on automation [1,2] but also related to more recent works on human-computer and human-robot interaction [3,4], the importance of trust in human-AI interactions has been theoretically and empirically established [5,6].

Moreover, with trust as a central construct in understanding human-AI interactions, it should be noted that trust is the subjective perception of AI's trustworthiness which does not necessarily reflect AI's actual trustworthiness. Such misalignments of perceived and actual trustworthiness can be the result of various biases. While some sources of such biases stem from specific design features [7], in this study, we are interested in biases originating in the human cognitive system. To that end, studies from cognitive and social psychology have identified a wide range of cognitive biases which systematically alter human perceptions of the world, such as availability bias, representativeness heuristic, or anchoring bias. Moreover, for social interactions it was found that human perceptions oftentimes aim to protect a person's (social) identity. One such social source of bias is motivated reasoning which suggests that individuals sometimes process information in a way to protect and confirm prior beliefs [8] and identities [9,10].

While motivated reasoning is well established in human-human interactions, in this study, we want to explore whether we can assume similar patterns when humans interact with AI. To that end, our main hypothesis is that receiving information from AI should reduce motivated reasoning as AI is oftentimes perceived to be less ideologically biased and more objective than its human counterparts [11,12]. Hence, our central research question is:

- RQ1: Can AI reduce motivated reasoning?

To answer this question, we conducted an online experiment in which we presented participants with polarizing information which was allegedly either written by a human journalist or a news algorithm, and asked participants to evaluate the trustworthiness of the information. Contextually, we locate our work in the field of journalism where, over the past years, AI systems have increasingly been employed as part of, for example, recommender systems and personalization, newsbots, the structuring and collection of data but also in automated storytelling [13].

2. Theoretical background

2.1. From trust in human agents to trust in AI

A central role in analyzing the interaction between humans and AI plays the notion of trust. However, trust as a multidimensional construct ([14]) comes with many definitions (see, e.g., [15,16]). Most of such definitions of trust comprise two central entities, (1) the trustor who places trust in someone (or something) and (2) the trustee who receives trust to perform a specific task. In addition, there must be the possibility that the trustee intentionally or unintentionally fails to perform the task, introducing the notion of risk, vulnerability and uncertainty to the interaction of trustor and trustee [1]. Such definitions of trust, originating in works on interpersonal trust between two human agents, have been applied to describe how humans relate to automated systems [1,5], robots [3] and AI [7].

Importantly, trust is the result of AI's subjectively perceived trustworthiness. Consequently, while AI systems can be more or less trustworthy due to their func-

tionality and reliability, users might *perceive* these systems differently. Ideally, a system's trustworthiness and its perceived trustworthiness can (or should) be related which has been labeled warranted [17] or calibrated trust [1]. In some situations, however, perceptions of trustworthiness do not reflect a system's actual trustworthiness, resulting in trustworthy systems being distrusted [18] and untrustworthy systems being trusted, leading to unwarranted trust [17] and misuse [18].

Consequently, it is important to identify which factors affect individuals' trust perceptions. To that end, especially research from social and cognitive psychology has shown that individuals' perceptions can be biased. While such biased perceptions can be the result of specific system features like anthropomorphism which has been found to increase the perceived trustworthiness of a system [7] [1], in this study, we are interested in biases originating in the human cognitive system.

To that end, two important factors, which have been found to systematically affect trust perceptions are (1) attitudes towards the trustee (in our case AI) [19,20] and (2) motivated reasoning [21]. In the following, we discuss how both are likely to affect trust in AI individually and jointly.

2.2. Attitudes towards AI

Two prominent and related perspectives which systematically describe individuals' trustworthiness perceptions of AI are algorithm appreciation/aversion and the machine heuristic. Works on *algorithm appreciation*, for example, have found that individuals prefer algorithmic advice over human advice (e.g., [12]). However, this preference has also been observed in the opposite direction, labeled as *algorithm aversion*. Untangling under which circumstances we can expect either algorithm appreciation or algorithm aversion, [22] investigated the role of framing effects and perceived expertise on trust. As a key result, the authors found that framing an agent to be perceived as higher in expert power explained whether individuals preferred algorithmic, or human advise.

Findings by [22] are well in line with predictions by the so-called *machine heuristic*. Building on earlier works, in which it was found that individuals preferred computer-generated information over human-generated information [23], [11] introduced the Modality-Agency-Interactivity-Navigability model (MAIN model). Central to this model is the assumption that individuals employ cognitive rules of thumb, that is cognitive heuristics, to navigate information systems such as the Internet. Because of the sheer information overload that the Internet offers, such heuristics allow individuals to quickly and effortlessly arrive at conclusions about information's trustworthiness. One of the central heuristics introduced by Sundar [11] is the machine heuristic. According to Sundar [11], the machine heuristic is triggered whenever specific cues suggest that individuals are dealing with a machine (in our case AI) rather than with a human. As a consequence, individuals apply common epistemic assumptions associated with ma-

[1] We understand anthropomorphism as a factor eliciting unwarranted trust, as anthropomorphism in itself does not increase the system's actual trustworthiness in terms of performance but merely increases its perceived performance

chines such as being ideologically unbiased, and being objective, which, in turn, lead to higher trust ratings [23].

Following the logic of the machine heuristic, we hypothesize that:

H1: For individuals who perceive AI as more competent and objective and less ideologically biased (higher trait belief in the machine heuristic), AI authorship should result in higher levels of perceived trust (as compared to human authors).

2.3. Motivated reasoning: Effects of opinion congruency

Especially in communication studies, one important factor affecting trust perceptions is opinion congruency, wherein opinion congruent sources are trusted more than opinion incongruent sources [24]. One theory that explains these dynamics is motivated reasoning. Motivated reasoning, a theory from social psychology [8], suggests that individuals sometimes process information in a biased manner, leading to overcritical assessment and rejection of opinion incongruent sources and information as well as lacking scrutiny and quick acceptance of opinion congruent sources and information. Empirically, motivated reasoning is well established in various contexts, but can we expect similar mechanisms when humans interact with AI?

To that end, first results on congruency effects have already been observed. For example, [25] found that individuals trusted an automated security officer (ASO) less when the judgment of the ASO differed from their own judgement. Similarly, [26] found that teachers accepted an AI-based educational technology less when it was incongruent to their own perception of a student and [27] found that . Furthermore, [27] investigated how physicians were more likely to reject a diagnosis from automated decision-support systems when diagnosis did not match their own.

While these results are insightful insofar as they provide evidence for congruency effects, the study designs do not include a human comparative condition. In other words, the studies do not illuminate whether the effects of congruency are unique to AI or whether we can expect a preference for *any* (AI and human) actor who shows congruency. Overcoming this limitation, [28] found that congruency increased adherence to AI-assisted recommendations but to a similar degree as recommendations by a co-worker. Hence, [28] could show that individuals did not differentiate between AI and a co-worker in terms of motivated reasoning. Receiving information from both, AI and the co-worker, resulted in the same pattern of motivated reasoning.

In contrast, [29] found that news written by AI were less likely to induce selective exposure, a concept which has been linked to motivated reasoning [30]. Similarly, [31] found that AI generated news were perceived as less biased (independent of opinion congruency), leading the author to suspect that "this finding offers possible optimism that the introduction of automation services may be capable of playing a role in reducing reactance to news perceived as biased" (p. 93).

Building on these results, we argue that, unlike humans as inherently agentic agents who can follow their own subjective agenda, AI systems should be perceived as less agentic and more objective, leading to less rejection of opinion-incongruent information. Hence, we hypothesize that:

H2: AI authorship should result in higher levels of perceived trust (as compared to human authors) when individuals are presented with information that is incongruent to their prior opinion.

Lastly, we connect predictions related to the machine heuristic and motivated reasoning. We expect that the effect suggested in H2 can partly be explained by the machine heuristic which proposes that individuals perceive machines as less biased than humans (see previous section). In other words, we hypothesize that:

H3: Individuals who perceive AI as more competent and objective and less ideologically biased should perceived incongruent information as more trustworthy when they are attributed to an AI author as compared to a human author (moderated moderation).

2.4. Setting the context: AI systems in journalism

Especially in the context of journalism, the specific perceived trustworthiness of information is commonly referred to as the credibility of information [32] - which we will follow in this work. Moreover, locating our work in the field of journalism, we built on previous studies which compare credibility perceptions of news generated by humans with news generated by AI (e.g., [33,34,35]). To that end, while results from some studies found that AI was trusted more than a human journalist [23], others found that individuals preferred human authors [35]. This is reflected in results from a recent meta-analysis by [36] who found, analyzing 12 studies which compared human with AI generated content, no clear preference for one over the other. Hence, we can assume that other factors affect how individuals place trust in AI systems versus human authors, such as expert framing [22]. In this work, we will explore the possible interaction of how an AI system is perceived (general preference of AI over humans) and what is communicated (congruent versus incongruent information), a connection which has not been made.

Moreover, vulnerability and risk are prerequisites of trust. Without vulnerability there is no need for trust as any none-performance would not affect the trustor. While it might not be straight forward to think that reading the news induces some kind of vulnerability, we argue that this is indeed the case, to the effect that individuals have to assess the credibility of information if they want to avoid being misinformed. In turn, being misinformed can have detrimental consequences to the individual concerning, for example, the individuals' health [37].

3. Method

The study received ethical approval from the ethics committee of the University of Duisburg-Essen. All hypotheses, the study design, and data analyses were preregistered via OSF. The data and data analysis syntax can be found here.

3.1. Sample, experimental design and procedure

We recruited 477 participants (255 female, 238 male, 11 non-binary, 3 preferred not to say) via the online survey platform Prolific. The mean age of participants

was $M = 31.24$ ($SD = 10.38$) and ranged from 18 to 71 years. Indicating the highest degree received, 34 participants indicated to have received a middle school degree, 107 a high school degree, 62 completed an apprenticeship, 259 a university degree (Bachelor or Masters), and 14 none of the above.

To test our hypothesis, we conducted a 2 x 2 between-subjects design with the independent factors, author and congruency. The factor *author* consisted of the two levels human author (n = 194) and AI author (n = 200). The factor *congruency* consisted of the two levels, opinion-congruent (n = 186) and opinion-incongruent (n = 208), and refers to the agreement between the participants' opinion and the opinion displayed in the news article. To determine congruency, participants were asked to indicate their opinion on the issue of gender-neutral language, which is either congruent or incongruent with the stance of the news article on gender-neutral language. Because motivated reasoning can only be expected when individuals hold an opinion for or against an issue, participants who did not express an opinion in support or against the issue were excluded from the analysis (for similar procedures, see [38,39]).

To begin with, participants were asked to provide standard demographic data. Subsequently, participants were introduced to the task. Depending on the experimental condition, participants were told that they were about to read a news article written either by a human journalist or by a news algorithm. When presented with the news article on gender neutral language, the author was repeated in the article's byline. After reading, participants were asked to evaluate the general message credibility. The study closed with participants indicating their attitude towards machines (machine heuristic), their prior experience with algorithms, and whether they remembered the author of the news article (manipulation check), after which participants were debriefed.

3.2. Stimulus material

The study participants were asked to read a short news article about the polarizing topic 'gender-neutral language in Germany'. The polarizing potential of the selected topic was tested in a pretest with N = 51 participants. Both articles were written by a human author and were written to represent supporting or contradicting arguments for gender-neutral language usage. The length of each article was between 236 and 247 words. To make sure possible differences in the perceived trustworthiness were not a result of the argument quality, text wording in both articles was kept as similar as possible by using negations.

3.3. Measures

We adopted [40]'s message credibility scale to assess the perceived trustworthiness of the news article. To that end, participants were asked to indicate how well the three adjectives can describe the news article: accurate, authentic, and believable (from 1 = describes very poorly to 7 = describes very well). All three items were summarized into one mean score with a Cronbach's $\alpha = .876$.

To assess individuals' stance on the machine heuristic, we adapted items created by [31]. In four items, measured on a 7-point Likert-type scale (from 1 =

Table 1. Unstandardized regression coefficients, standard errors, and p-values for the basic regression model (including only controls), model 1 (testing H1), and model 2 (testing H2).

	model 0			model 1			model 2			model 3		
	B	*SE*	*p*	B	*SE*	*p*	B	*SE*	*p*	B	*SE*	*p*
congruency	-1.16	0.12	.001	-1.07	0.12	.001	-0.10	0.39	.793	-0.10	0.38	.796
author	0.09	0.12	.465	0.95	0.37	.010	2.05	0.56	.001	1.05	0.36	.003
age	-0.01	0.01	.356	-0.01	0.01	.441	-0.01	0.01	.682	0.01	0.01	.738
gender	-0.20	0.11	.072	-0.16	0.11	.136	-0.16	0.11	.139	-0.15	0.11	0.16
AI experience	0.09	0.08	.273	0.11	0.08	.184	0.12	0.08	.137	0.13	0.08	.138
MH				0.54	0.14	.001	0.60	0.14	.001	0.22	0.45	.649
author*MH				-0.22	0.09	.014	-0.26	0.09	.005	-0.10	0.31	.744
author*congruency							-0.64	0.24	.008	-0.65	0.24	.008
author*congruency*MH										-0.10	0.21	.637

Congruency is coded 1 = congruent, 2 = incongruent; author is coded 1 = AI, 2 = human. MH is short for machine heuristic.

strongly disagree to 7 = strongly agree), participants were asked to indicate their agreement the following statements: (1) if a machine does a job, then the task was done objectively, (2) if a machine does a job, then the work was error-free, (3) if a machine does a job, then the work was unbiased, and (4) if a machine does a job, then the task was done accurately. For the analysis, items were summarized into one mean score with Cronbach's $\alpha = .797$. The mean perceived machine agency was 3.79 ($SD = 1.30$).

Besides the standard demographic data described above, we also controlled for algorithmic expertise. Participants were asked to indicate their prior knowledge about AI as follows: "A great deal/I am an expert on this topic," (n = 25), "A lot/I have read quite a lot about," (n = 116), "Some/I have some knowledge," (n = 269), "A little/I have only heard of artificial intelligence," (n = 66) or "None/I have no idea what artificial intelligence is." (n = 1).

4. Results

4.1. Hypotheses testing

To test our hypotheses, we ran four multiple linear regressions (for coefficients, standard errors, and p-values of all models, see Table 1). In model 0, we included only the control variables age and gender. Since the main effects of author and congruency were not of prior interest, we included both in this basic model as well. Results of the multiple regression analysis could show that model 0 predicted the perceived trustworthiness of the article significantly better than the null model (the mean) with $F(6, 387) = 16.67$, $p < .001$, $R^2 = .193$. Moreover, we found that if participants' prior opinion was congruent to the article's content, perceived trustworthiness increased, supporting previous results of motivated reasoning [41]. In addition, we also found that participants perceived the text purportedly written by AI similar to the text purportedly written by a human author.

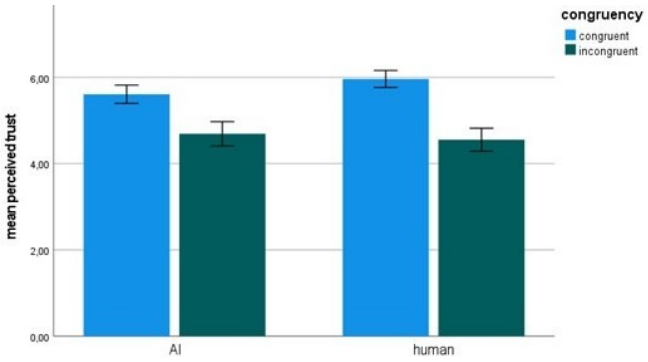

Figure 1. Perceived trustworthiness by condition.

4.1.1. Hypothesis 1

Following this basic model, we successively added predictors according to our hypotheses in model 1, model 2, and model 3. In H1, we hypothesized that, for individuals who perceive AI as more competent and objective and less ideologically biased (higher trait belief in the machine heuristic), the AI author should result in higher levels of perceived trust (compared to the human author). Similar to model 0, model 1 was significant with $F(8, 385) = 15.68$, $p < .001$, $R^2 = .23$. Moreover, the model was significantly better than the null model ($F(2, 385) = 10.3$, $p < .001$).

Our results support H1. Participants' belief in the machine heuristic affect how participants perceived the AI author compared to the human author. Participants who perceive AI as more competent, objective and less ideologically biased perceived the news article as more trustworthy when it was purportedly written by AI than by a human. Interestingly, when including the effects of belief in the machine heuristic, the effect of author became significant, implying that the authors, AI and human, were indeed perceived differently. The text purportedly written by a human was perceived more trustworthy than the text puportedly written by AI.

4.1.2. Hypothesis 2

In H2, we predict that AI authorship should result in higher levels of perceived trust (as compared to human authors) when individuals are presented with information that is incongruent to their prior opinion. Model 2 was significant with $F(9, 384) = 14.91$, $p < .001$, $R^2 = .24$, and significantly better than model 1 ($F(1, 384) = 6.84$, $p = .009$). Supporting H2, we found that the interaction of authorship and congruency was also significant. We visualized the results in Figure 1.

Inspecting the relationship of author and congruency, we found that, contradictory to our predictions, the effect was due to reduced trust in the congruent condition ($t(184) = 2.42$, $p = .016$). There was no difference between AI and human authorship when the article was incongruent to the participants' opinion ($t(206) = 0.69$, $p = .448$).

To better understand this effect, we wanted to know whether the observed difference between trust perceptions was driven by participants overtrusting opinion congruent content or participants placing less trust in opinion incongruent content. While, in this study, we did not include a control condition (or baseline condition) per se, we could compare opinion congruent and incongruent trust perceptions with those perceptions by participants who indicated to hold no directional attitude towards gender-neutral language. Through a one-way analysis of variance (ANOVA) and subsequent post-host test (with a Bonferroni corrected alpha level of α =.0125), we found that the observed difference was indeed driven by participants overtrusting opinion congruent content (independent of the author), whereas participants' perceptions of trust in the opinion incongruent condition did not significantly differ from those who held no prior opinion (control group vs incongruent$_{\text{AI}}$: $p = .822$, 99.2% CI = $[-0.43, 0.84]$; control group vs incongruent$_{\text{human}}$: $p = .356$, 99.2% CI = $[-0.28, 0.98]$[2]).

4.1.3. Hypothesis 3

In H3, we hypothesized that individuals who perceive AI as more competent and objective and less ideologically biased should perceive incongruent information as more trustworthy when attributed to an AI author than a human author. To test this moderated moderation hypothesis in model 3, we used Process for R (version 4.0.1) by [42]. The resulting model 3 was significant with $F(11, 382) = 11.89$, $p < .001$, $R^2 = .26$. Results indicated, however, a non-significant effect. Consequently, H3 had to be rejected.

5. Discussion

In this work, we investigated which biases systematically affect trust perceptions and how they play out. While AI is entrusted with various tasks, we situated our investigation within the realm of AI journalism, comparing how individuals perceive purportedly human-generated content compared to purportedly AI-generated content. To that end, building on insights related to the machine heuristic and algorithm appreciation, both of which suggest that individuals perceive machines/AI as more competent and objective and less ideologically biased, we hypothesized that the higher individuals' trait belief in the machine heuristic is, the more they would perceive content produced by AI as more trustworthy than a human author.

Our results support this hypothesis. Individuals who indicated that AI is more competent and objective and less ideologically biased than humans perceived AI news as more trustworthy than human news. These results are in line with previous findings, which could show, for example, that individuals are more willing to disclose information to a machine than to another person if they thought that machines were more trustworthy than humans [43].

Interestingly, while we initially did not find a difference in perceived trust between purportedly human and purportedly AI generated text, this changed when

[2]All means and standard deviations are reported in the supplement material S1.

we included participants' attitudes towards AI in the model. Filtering the effect of attitudes towards AI revealed that participants in our study found content from the human author more trustworthy than the AI author. These results echo previous findings, showing that individuals trust human authors more than AI authors (e.g., [31]), but also stress how attitudes towards AI shape these perceptions.

In addition to the effects of attitudes towards AI, we were also interested in possible effects of motivated reasoning. To that end, literature on motivated reasoning (e.g., [8]) suggests that individuals are overly critical of and more likely to distrust information that does not support their prior opinion, whereas information that supports their prior opinion is more readily assimilated trusted, and believed. Building on this, we predicted that AI authorship should result in higher levels of perceived trust (compared to human authors) when individuals are presented with information incongruent with their prior opinion.

As predicted, our results indicate that authorship (AI versus human) and the congruency condition interacted. This implies that the perceived trustworthiness of an AI author and a human author, respectively, depended on whether or not the article was (in)congruent to the individual's prior opinion. Inspecting this interaction, however, we found that, unlike predicted, the AI authorship reduced the effects of opinion congruency and not incongruency. To be precise, when reading the seemingly AI-generated content, participants showed less bias for opinion-congruent content. In contrast, opinion-incongruent content remained unaffected by the authorship condition.

Why did we find this? We explain this unexpected finding by referring to the main effect we found for authorship. As reported above, when filtering out the effects of attitudes towards AI, we found that participants trusted the content by the human author more than the (in fact, same) content by AI. We assume that this lack of trust towards AI is reflected in the reduced trust perceptions even when content affirmed participants' opinion. Hence, while our original assumption suggests that AI would increase trust perceptions for opinion incongruent content, we found that AI could reduce biased trust perceptions by reducing overly trustworthy perceptions of due to opinion congruent content.

To some degree, our results align with previous findings from [29], who found that participants of their study preferred opinion congruent news written by a human author over opinion congruent news written by an AI. The authors explain their findings pointing to increased levels of perceived trust towards human news compared to AI news.

In our last hypothesis, we wanted to know whether attitudes towards AI could explain the relationship between opinion-congruency and authorship. However, we did not find any significant interactions (moderated moderation).

5.1. Implications

What becomes apparent through our work is that specific attitudes towards AI and ideological biases due to motivated reasoning systematically shaped trust perceptions. Various theoretical and practical implications arise from these takeaway findings that go beyond our work's contextual scope (AI in journalism).

First, thinking about trustworthy AI, our work implies that, to understand trust perceptions, we need to consider how individuals understand, make sense,

and perceive AI. More specifically, our work shows that, while an AI author was generally perceived as less trustworthy than a human author, the exact opposite effect was found for individuals who perceive AI as more competent and objective and less ideologically biased. Hence, disregarding individuals' attitudes would have wrongly resulted in a null finding, hiding the underlying dynamics. Knowing how end-users perceive AI is, hence, essential.

To that end, various studies have measured attitudes towards AI. In a large-scale survey, for example, [44] have asked over 150,000 respondents in 142 countries about potential harms and benefits of AI. However, knowing about these attitudes is just one step towards trustworthy AI. We also need to ask how individuals arrive at such attitudes and how attitudes might be changed. [43] suggest that one source of such attitudes are, for example, "common stereotypes about machines" (p. 2). However, the authors do not explain how such stereotypes evolve. Another perspective comes from [22], who suggest that perceptions of AI depend on how AI is framed in terms of its competence/expert power by introducing an AI system either as above-average performing or as average performing. These results also align with previous findings by [45], who found that depending on the terminology used to describe AI, for example, as a computer program, an automated system, machine learning, or an algorithm, affected how participants perceived AI in terms of, for example, system complexity, fairness, and trust. Hence, our results, combined with previous findings, stress the importance of attitudes towards AI and how these must be addressed when designing trustworthy AI systems.

Second, similarly to the effects of attitudes towards AI, our work shows that trust perceptions are shaped by individuals' prior opinion towards the task outcome. While previous studies could already show that an AI author led to reduced hostile media perceptions [46,47] and reduced perception of media bias [31], our study adds to this that AI could reduce effects of motivated reasoning. In other words, when the task outcome (the news article) was congruent to the participants' opinion, participants trusted it more than participants who did not have a prior opinion or those whose opinion was not supported. And although this effect was reduced in the AI condition compared to the human condition, levels of trust were significantly higher in the congruent condition compared to the incongruent and neutral condition, indicating a case of overtrust.

Moreover, we found that our findings are the results of specific attitudes towards AI, similar to previous outcomes. For example, the reduced hostile media perception [46] and the reduced media bias [31] were the result of perceived superiority of AI compared to humans in terms of competence and objectivity (increased belief in the machine heuristic). In turn, our findings were the result of reduced trustworthiness of AI. Hence, attitudes towards AI also shape how biases like motivated reasoning occur. However, it remains open how such dynamics change when attitudes towards AI change. For example, can we expect that perceived media bias decreases when perceptions about AI turn from competent and objective to biased? For now, belief in the machine heuristic and automation biases remain, as [43] suggested, due to common stereotypes, but how would those stereotypes change as the public becomes more aware of algorithmic biases, a concept widely discussed and acknowledged in academia (e.g., [48,49]).

To conclude, both findings, the effects of attitudes towards AI and motivated reasoning, are significant because they stress the importance of how we approach and measure trustworthy AI. In our case, if we had simply asked participants to indicate how trustworthy they perceived the news article, we would have found no difference between the human author and the AI author. Only when including the joint effects of attitudes towards AI and the congruency information, did we find that trust perceptions indeed varied. When including participants' attitudes (in this case, the belief in the machine heuristic) and their prior opinion, we could excerpt more fine-grained effects that were otherwise hidden.

5.2. Limitations

In the light of the discussed implications of our work, we would like to point to equally important limitations. First, as pointed out above, we neither anticipated nor know why the human author was perceived as more trustworthy than the AI author. Previous studies suggest that such an effect might be related to anthropomorphism [31] or emotional involvement [46], but since we did not include any such measure or manipulations into our study, we cannot explain this result.

Second, our study is restricted to trait assumptions, that is, participants' stable beliefs in the machine heuristic. To increase the causal claim that these beliefs/attitudes towards AI shape trust perceptions, future studies should include manipulations of such, similar to [22].

Third, given the extensive range of AI applications (e.g., recommender systems, decision aids, voice assistants, or autonomous vehicles) that vary in their task affordances, we are limited to the task we used, automated content creation. To that end, automated content creation neither relies on user input like recommender systems or voice assistance nor is the task characterized by increased levels of risk such as some decision aids or autonomous vehicles. Hence, the generalizability of our findings to other tasks is inconclusive.

6. Conclusion

In this work, we asked whether AI could debias individuals' trust perceptions. In addition, we were interested in how attitudes about AI might affect such (de)biased trust perceptions. To answer these questions, we asked participants to read either a purportedly AI or purportedly human-generated news article and asked them how trustworthy they found the content.

Our results indicated that (a) attitudes towards AI shaped trustworthiness perceptions and that (b) AI debiased trust perceptions in cases where participants were swayed by the prior opinions to trust content more when they agreed with the content. Notably, the difference in trust perceptions of the AI author versus the human author only became apparent when accounting for attitudes towards AI and prior opinions, emphasizing the explanatory importance of attitude and prior opinions.

References

[1] Lee JD, See KA. Trust in automation: Designing for appropriate reliance human factors. Human Factors. 2004;46(1):50-80. Available from: http://csel.eng.ohio-state.edu/productions/intel/research/trust/Lee&SeeTrustReview.pdf.

[2] Madhavan P, Wiegmann DA. Similarities and differences between human–human and human–automation trust: An integrative review. Theoretical Issues in Ergonomics Science. 2007;8(4):277-301.

[3] Hancock PA, Billings DR, Schaefer KE, Chen JYC, De Visser EJ, Parasuraman R. A meta-analysis of factors affecting trust in human-robot interaction. Human Factors. 2011;53(5):517-27.

[4] Sanders T, Oleson KE, Billings DR, Chen JYC, Hancock PA. A model of human-robot trust. Proceedings of the Human Factors and Ergonomics Society. 2011:1432-6. Available from: http://dx.doi.org/10.1177/1071181311551298.

[5] Hoff KA, Bashir M. Trust in automation: Integrating empirical evidence on factors that influence trust. Human Factors. 2015;57(3):407-34.

[6] Oksanen A, Savela N, Latikka R, Koivula A. Trust toward robots and artificial intelligence: An experimental approach to human–technology interactions online. Frontiers in Psychology. 2020;11:1-13.

[7] Kaplan AD, Kessler TT, Brill JC, Hancock PA. Trust in artificial intelligence: Meta-analytic findings. Human Factors. 2021. Available from: https://doi.org/10.1177/00187208211013988.

[8] Kunda Z. The case for motivated reasoning. Psychological Bulletin. 1990;108(3):480-98. Available from: http://doi.apa.org/getdoi.cfm?doi=10.1037/0033-2909.108.3.480.

[9] Kahan DM. Misconceptions, misinformation, and the logic of identity-protective cognition; 2017. Available from: https://ssrn.com/abstract=2973067.

[10] Van Bavel JJ, Pereira A. The partisan brain: An identity-based model of political belief. Trends in Cognitive Sciences. 2018;22(3):213-24.

[11] Sundar SS. The MAIN model: A heuristic approach to understanding technology effects on credibility. Digital Media, Youth, and Credibility. 2008:73-100. Available from: http://www.mitpressjournals.org/doi/abs/10.1162/dmal.9780262562324.073.

[12] Logg JM, Minson JA, Moore DA. Algorithm appreciation: People prefer algorithmic to human judgment. Organizational Behavior and Human Decision Processes. 2019;151(April 2018):90-103. Available from: https://doi.org/10.1016/j.obhdp.2018.12.005.

[13] Broussard M, Diakopoulos N, Guzman AL, Abebe R, Dupagne M, Chuan CH. Artificial intelligence and journalism. Journalism & Mass Communication Quarterly. 2019;96(3):673-95.

[14] Ullman D, Malle BF. What does it mean to trust a robot? Steps toward a multidimensional measure of trust. In: Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction; 2018. p. 263-4.

[15] PytlikZillig LM, Kimbrough CD. Consensus on conceptualizations and definitions of trust: Are we there yet? In: Interdisciplinary Perspectives on Trust. Springer; 2016. p. 17-47.

[16] Mcknight DH, Carter M, Thatcher JB, Clay PF. Trust in a specific technology: An investigation of its components and measures. ACM Transactions on Management Information Systems. 2011;2(2):1-25.

[17] Jacovi A, Marasović A, Miller T, Goldberg Y. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. FAccT 2021 - Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. 2021;(Section 2):624-35.

[18] Parasuraman R, Riley V. Humans and automation: Use, misuse, disuse, abuse. Human Factors. 1997;39(2):230-53.

[19] Mayer RC, Davis JH, Schoorman FD. An integrative model of organizational trust. Academy of Management Review. 1995;20(3):709-34.

[20] Davis FD, Bagozzi RP, Warshaw PR. User acceptance of computer technology: A comparison of two theoretical models. Management Science. 1989;35(8):982-1003.

[21] Freiling I, Waldherr A. Why Trusting Whom? Motivated Reasoning and Trust in the Process of Information Evaluation. In: Trust and Communication. Springer; 2021. p. 83-97.

[22] Hou YTY, Jung MF. Who is the expert? Reconciling algorithm aversion and algorithm appreciation in AI-supported decision making. Proceedings of the ACM on Human-Computer Interaction. 2021;5(CSCW2).
[23] Sundar SS, Nass C. Conceptualizing sources in online news. Journal of Communication. 2001;51(1):52-72.
[24] Westerwick A, Johnson BK, Knobloch-Westerwick S. Confirmation biases in selective exposure to political online information: Source bias vs. content bias. Communication Monographs. 2017;84(3):343-64.
[25] Huang HY, Twidale M, Bashir M. 'If you agree with me, do i trust you?': An examination of human-agent trust from a psychological perspective. vol. 1038. Springer International Publishing; 2020. Available from: http://dx.doi.org/10.1007/978-3-030-29513-4_73.
[26] Nazaretsky T, Cukurova M, Ariely M, Alexandron G. Confirmation bias and trust: Human factors that influence teachers' attitudes towards AI-based educational technology. CEUR Workshop Proceedings. 2021;3042.
[27] Dreiseitl S, Binder M. Do physicians value decision support? A look at the effect of decision support systems on physician opinion. Artificial Intelligence in Medicine. 2005;33(1):25-30.
[28] Alon-Barkat S, Busuioc M. Human-AI interactions in public sector decision-making: 'Automation bias' and 'selective adherence' to algorithmic advice. Journal of Public Administration Research and Theory. 2022;muac007.
[29] Jia C, Johnson TJ. Source credibility matters: Does automated journalism inspire selective exposure? International Journal of Communication. 2021;15:22.
[30] Yeo SK, Cacciatore MA, Scheufele DA. News selectivity and beyond: Motivated reasoning in a changing media environment. In: Jandura O, Petersen T, Schielicke A, editors. Publizistik und gesellschaftliche Verantwortung. Wiesbaden: Springer Fachmedien; 2015. p. 83-104.
[31] Waddell TF. Can an algorithm reduce the perceived bias of news? Testing the effect of machine attribution on news readers' evaluations of bias, anthropomorphism, and credibility. Journalism and Mass Communication Quarterly. 2019;96(1):82-100.
[32] Metzger MJ, Flanagin AJ. Credibility and trust of information in online environments: The use of cognitive heuristics. Journal of Pragmatics. 2013;59:210-20. Available from: http://dx.doi.org/10.1016/j.pragma.2013.07.012.
[33] Clerwall C. Enter the robot journalist: Users' perceptions of automated content. Journalism Practice. 2014;8(5):519-31.
[34] Jung J, Song H, Kim Y, Im H, Oh S. Intrusion of software robots into journalism: The public's and journalists' perceptions of news written by algorithms and human journalists. Computers in Human Behavior. 2017;71:291-8.
[35] Waddell TF. A robot wrote this? How perceived machine authorship affects news credibility. Digital Journalism. 2018;6(2):236-55.
[36] Graefe A, Bohlken N. Automated journalism: A meta-analysis of readers' perceptions of human-written in comparison to automated news. Media and Communication. 2020;8(3):50-9.
[37] Suarez-Lledo V, Alvarez-Galvez J. Prevalence of health misinformation on social media: Systematic review. Journal of Medical Internet Research. 2021;23(1).
[38] Brenes-Peralta C, Wojcieszak M, Lelkes Y. Can I stick to my guns? Motivated reasoning and biased processing of balanced political information. Communication & Society. 2021;34(2):49-66.
[39] Drummond C, Fischhoff B. Does "putting on your thinking cap" reduce myside bias in evaluation of scientific evidence? Thinking and Reasoning. 2019;25(4):477-505. Available from: https://doi.org/10.1080/13546783.2018.1548379.
[40] Appelman A, Sundar SS. Measuring message credibility: Construction and validation of an exclusive scale. Journalism & Mass Communication Quarterly. 2016;93(1):59-79.
[41] Wischnewski M, Krämer N. I reason who I am? Identity salience manipulation to reduce motivated reasoning in news consumption. In: Proceedings of the 11th International Conference on Social Media and Society; 2020. p. 1448-155.
[42] Hayes AF. Introduction to mediation, moderation, and conditional process analysis: A regression-based approach. Guilford publications; 2017.

[43] Shyam Sundar S, Kim J. Machine heuristic: When we trust computers more than humans with our personal information. Conference on Human Factors in Computing Systems - Proceedings. 2019:1-9.

[44] Neudert LM, Knuutila A, Howard PN. Global Attitudes Towards AI, Machine Learning & Automated Decision Making. Oxford Internet Institute; 2020.

[45] Langer M, Hunsicker T, Feldkamp T, König CJ, Grgić-Hlača N. "Look! It's a computer program! It's an algorithm! It's AI!": Does terminology affect human perceptions and evaluations of intelligent systems? 2021;1(1):1-28. Available from: `http://arxiv.org/abs/2108.11486`.

[46] Liu B, Wei L. Reading machine-written news: Effect of machine heuristic and novelty on hostile media perception. vol. 10901. Springer International Publishing; 2018. Available from: `http://link.springer.com/10.1007/978-3-319-91238-7`.

[47] Cloudy J, Banks J, Bowman ND. The str(AI)ght scoop: Artificial intelligence cues reduce perceptions of hostile media bias. Digital Journalism. 2021:1-20.

[48] Diakopoulos N. Algorithmic accountability: Journalistic investigation of computational power structures. Digital journalism. 2015;3(3):398-415.

[49] Danks D, London AJ. Algorithmic bias in autonomous systems. In: IJCAI. vol. 17; 2017. p. 4691-7.