


We and It: An Interdisciplinary Review of the Experimental Evidence on Human-Machine Interaction

Working Paper

Author(s):

Chugunova, Marina; [Sele, Daniela](#) 

Publication date:

2020-08

Permanent link:

<https://doi.org/10.3929/ethz-b-000442053>

Rights / license:

[In Copyright - Non-Commercial Use Permitted](#)

Originally published in:

Center for Law & Economics Working Paper Series 12/2020

Center for Law & Economics Working Paper Series

Number 12/2020

**We and It: An Interdisciplinary Review of the
Experimental Evidence on Human-Machine Interaction**

**Marina Chugunova
Daniela Sele**

August 2020

We and It: An interdisciplinary review of the experimental evidence on human-machine interaction *

Marina Chugunova [†] Daniela Sele [‡]

August 21, 2020

Abstract

Today, humans interact with technology frequently and in a variety of settings. Their behavior in these interactions has attracted considerable research interest across several fields, with sometimes little exchange among them and seemingly inconsistent findings. Here, we review over 110 experimental studies on human-machine interaction. We synthesize the evidence from different disciplines, suggest ways to reconcile inconsistencies, and elaborate on political and societal implications. The reviewed studies show that people react to automated agents differently than to humans: They behave more rationally, and are less prone to emotional and social responses. We show that there are several factors which systematically impact the willingness to accept automated decisions: task context, performance expectations and the distribution of decision authority. That is, humans seem willing to (over-)rely on algorithmic support, yet averse to fully ceding their decision authority. These behavioral regularities need to be considered when deliberating the benefits and risks of automation.

Keywords: human-computer interaction, human-machine interaction, algorithmic decision making, experimental evidence, review.

JEL-classification: O33, C90, D90.

*We thank Elliott Ash, Stefan Bechtold, Dietmar Harhoff, Wolfgang Luhan, Christopher Sprigman and Kai-Uwe Schnapp for their insightful comments, as well as Nishan Lin, Sarah Doelger and Damjan Kostovic for valuable research assistance.

[†]Max Planck Institute for Innovation and Competition, 80539 Munich, Germany.

[‡]Center for Law & Economics, ETH Zurich, 8092 Zurich, Switzerland.

1 Introduction

Early in the morning, James drives to work on the fastest route picked out for him by his navigation system. On the highway, his car's blind spot assistant helps him avoid a potentially dangerous situation. He happily listens to a playlist of new songs selected for him by an app. An alert on his phone warns him about unusual activity on his bank account, but the issue is swiftly resolved with the help of the bank's chat bot. After passing the automated barrier to enter his office, James plugs his portable computer into the docking station on his desk, and starts his day.

While James's morning may seem like nothing out of the ordinary, it serves to illustrate something rather extraordinary: how frequent and how routine interactions between humans and automated agents have become. Had someone read this paragraph a few decades ago, they would probably have assumed that our hero was James Bond. In contrast, today, automated agents are everywhere. They collaborate with humans in the workplace and trade with them on the stock market. They are involved in deciding whether a human should receive a job or be released from jail. And, recently, the move into a far more digital way of working and communicating caused by the global pandemic in 2020 has only accelerated the involvement of automated agents in daily life and decision-making.

What happens when interactions between humans turn into interactions between humans and automated agents? How do humans react to the presence of increasingly capable automated agents in the workplace, and to their increasing involvement in decision-making? The ubiquity of automation in today's world calls for an interdisciplinary effort to answer these questions (1). And, indeed, many researchers have investigated them. However, the fast-growing number of studies addressing these questions, the wide range of methodologies used as well as a number of (seemingly) inconsistent findings make it difficult for interested researchers or policy-makers to gain an overview of the existing evidence. To change this, we review the findings of 118 experimental studies that investigate how humans interact with automated agents.

The review covers studies from a range of disciplines, including psychology, economics, sociology, human-computer interaction, judgement and decision making, neuroscience, marketing and consumer research, computer science, information systems, medicine and even aeronautics. Due to the differences in terminology and keywords across disciplines, we primarily resorted to ancestry searching and journal hand searching to identify relevant articles (following (2)). We focus on experimental studies, with the aim to discuss causal effects as well as underlying mechanisms. As a result of the differences in methodology across fields, our sample of studies includes those that elicit stated preferences, revealed preferences or use physiological measurements, that use or don't use incentives, and that resort to various different subject pools. To guard against publication bias and to include the most recent findings we deliberately include publically available working papers (following (3)). The Appendix provides short notes on the methodologies of all cited experiments, observational studies and literature reviews on human-machine interaction.

We use a wide definition of human-computer interaction, and include papers studying the interaction of humans with automated agents, computers/computer systems, machines, robots, algorithms, AI systems etc. This allows us to draw on a large range of studies spanning from the early years of automation up until recent discussions of sophisticated machine learning algorithms. As we are interested in studying the behaviors and preferences humans exhibit when they interact with automated agents, we do not discuss research on the macro-economical consequences of automation, on the ethics of automation or on algorithm design. Similarly, as we are interested in uncovering patterns in behavior, we do not discuss specialized work on user experience (UX) or on interface design. Given the ambitious goal of the paper and the number of covered disciplines, the list of included papers is by no means exhaustive, but will hopefully provide a useful starting point to an interested reader.

The reviewed studies show that humans interact with automated agents in social ways, but differently than they do with fellow humans. They respond to automated agents' actions with less emotions, and are less (but not un-)concerned with social rules of conduct when interacting with automated agents. The type of task matters: humans seem willing to engage with automated agents in contexts perceived as analytical or objective, but reluctant to do so in more social or moral contexts. While studies of collaboration with autonomous automated agents in the workplace remain rare, the existing evidence shows that humans seem yet to develop effective ways to collaborate with automated agents in the workplace. Regarding the use of automated agents in decision-making, some studies show that humans are averse to delegating decision authority to automated agents - while others find them to be appreciative of automated advice, and sometimes even overreliant on automated decision-making supports. Finally, some studies find evidence that humans are willing to accept automated managers, in particular where the nature of the managerial decisions is perceived as analytical rather than social - though, again, further research seems to be needed. To enable such research, the review proposes concrete hypotheses to reconcile seemingly inconsistent findings.

The contribution of this review is threefold. First, in sections 2 and 3, we synthesize the evidence from numerous studies into a number of recurrent findings on how humans interact with automated agents. In section 3 on automated decision-making, we in addition scrutinize a number of seemingly inconsistent findings, and suggest a structured way to reconcile them that future research could investigate. This distinguishes us from other literature reviews, which tend to focus in more detail on a specific strand of the literature, leaving them unable to address inconsistencies between different fields (see e.g. (4) on algorithm aversion and appreciation, (5) on algorithm aversion, (6; 7) on automation bias and automation-induced complacency or (8) on the use of computer players in economic experiments. Brief descriptions of these and further reviews can be found in the Appendix.). Finally, in section 4, we link the empirical findings to the on-going political and societal discussion about the benefits and harms of automation. Hence, we discuss how behavioral research can inform effective regulation of automated decision-making. Finally, with all of this, we aim to promote the dissimilation of insights and ideas across disciplines, and to help provide policy-makers with grounds for evidence-based

decision-making.

2 Social interactions with automated agents

2.1 The perception of automated agents as social interaction partners

To start our discussion, let us first establish whether it is appropriate to consider interactions between humans and machines as social interactions. Researchers have long documented that humans tend to create narratives around events (9) and attribute agency to inanimate objects (10). In an early series of lab experiments in the aptly termed *computers are social actors* (CASA) paradigm, psychologists documented that people apply social rules and expectations to computers (for an overview see (11)): Participants applied gender stereotypes (12) and ethnic stereotypes (13) to computers. They reacted to automated feedback (14), preferred computers arbitrarily marked as team-mates (13) and reciprocated helpful acts by computers (15). Reactions of reciprocity were even adjusted to prevailing cultural norms (16). Yet, when asked directly, most participants in these experiments denied that they felt that computers had a personality, or that they warranted social or polite treatment (11).

Later studies provided further evidence for the social treatment of machines. People were again found to apply gender stereotypes (17; 18; 19) and racial stereotypes (20) to automated agents. People were also found to use social cues like smiles and silence fillers when interacting with automated agents (21), and to react to automated flattery (22). A recent experiment showed that robots can trigger actions of social conformity at a similar rate as humans - but only until participants lose trust in faulty robots (23). Given sufficient behavioral realism, social reactions can be triggered by both avatars (digital representations of humans) and virtual agents (interactive computer programs) (24).

It is important to note that these findings do not imply that automated agents are treated equally to humans in interactions. Indeed, people treat automated agents differently than other humans, as this review will discuss at length. These differences are visible in neurophysiological studies, which document that different areas of the brain are activated when humans interact with fellow humans or automated agents (see e.g. 25; 26; 27; 28). The reason behind the differential treatment of automated interaction partners might lie in the way humans perceive their respective counterparts' agency. Generally, in social interactions, people engage in the process of *mentalizing* to infer the mental state and capacities of their counterparts (29). This deliberation seems to occur less with automated agents: the area of the brain connected to mentalizing is activated less when humans interact with automated agents (28; 30). The reduced need to infer the mental state of one's counterpart might also explain the repeated finding that people react faster to actions of automated counterparts (see e.g. 30; 31). And, indeed, as a large survey shows, people do not attribute robots with minds (32).

Interestingly, this survey showed that robots are denied the capacity to experience moral right

or wrong (32). When robots are described as able to feel, people report feelings of unease (33). Despite this perception, however, people seem averse to mistreating automated agents. Indeed, people were found to exhibit both physiological and behavioral displays of stress when watching videos of robots being "tortured" (34), or when asked to administer "painful" electric shocks to robots (35) or virtual agents (36) in two Milgram-style experiments (37). Another experiment showed that people were unwilling to follow a command to destroy a tower a robot had built if the robot protested and started "sobbing" (38).

2.2 The reduced emotional and social response to automated agents

A particularly striking difference between human-human and human-computer interactions is the robust finding that interacting with automated agents triggers less of an emotional and social response in the human interaction partners. This phenomenon has been repeatedly documented using both subjective and behavioral measures, e.g. in trust games (39; 40) or in ultimatum, dictator and public goods games (41). It has also been found using physiological measurements, e.g. when playing computer games (42), in ultimatum games (27; 43) or in auctions (44; 45).

The reduced emotional response to automated agents is manifested both in terms of a less emotional immediate reaction to the agent's actions as well as in a generally decreased level of emotional arousal in human-computer interactions (as demonstrated, e.g., in (44)). Importantly, interacting with an automated counterpart seems to narrow the entire emotional spectrum: people react less positively to desirable actions by automated agents as well as less negatively to undesirable ones. For instance, in an experiment principals who delegated tasks reported significantly less enjoyment of good outcomes and significantly less anger for bad outcomes when the task was delegated to an algorithm rather than to a human (46). In a vignette study, observers rated a physician's correct medical decisions less positively and incorrect decisions less negatively if the physician was described to use an automated decision-making aid (47). In a social exchange experiment, participants perceived coercive actions by computers as less unjust than coercive actions by humans, and retaliated against them less (in contrast, here cooperative actions were perceived as similarly just regardless of who made them) (48). The reduced emotional response could be due to a perceived lack of intent on part of the automated agent (as speculated by (48), too).

An upside of this decreased emotional and social response is that it can increase rationality in interactions. For instance, the introduction of automated agents has been shown to reduce bubbles in simulated stock markets (31), or to increase bargaining efficiency in auctions (49). Computers can also help avoid undesirable social responses. For example, due to a reduction of social image concerns, participants were found more likely to disclose uncomfortable information to automated agents rather than to humans (50). Reporting to computers even significantly increased the likelihood of disclosures of intimate partner violence (51; 52).

On the downside, automation could prove harmful in contexts where beneficial behaviors are

driven by emotions and social concerns like pro-sociality, social comparisons or reciprocity. This is well demonstrated in an experiment in which participants play public goods, ultimatum and dictator games against other humans or against automated agents (41). In all three games, participants proved less willing to share with automated counterparts, and felt less guilty about exploiting them. This mirrors findings of earlier CASA studies, which documented participants' willingness to engage in self-serving bias with automated counterparts (i.e., to attribute positive outcomes to themselves and negative outcomes to the computer) (53; 54). The introduction of automated agents can also increase the willingness to engage in unethical behavior: When given a chance to misreport the outcome of a coin toss to increase their monetary profit, participants in an experiment were significantly more likely to lie when reporting to a computer (55). The lack of altruism and social pressure towards automated teammates was also shown to lower worker productivity in an experimental sequential assembly line task (56). In summary, while automation can have beneficial effects in contexts where emotions or social concerns are detrimental, it can also be harmful by reducing them in contexts where they are beneficial.

An important factor which affects the emotional and social response to an automated agent seems to be its behavior and appearance. Indeed, humans seem to prefer interacting with automated agents that behave similarly to humans, i.e. which display contingent verbal and non-verbal reactions (57) or relational behaviors (58). Unexpected behavior by robots can also amplify social responses: for example, people were more likely to display social reactions towards robots that unexpectedly cheated in a rock-paper-scissors game (59). In a study with service robots who delivered medicine to elderly patients, the use of human-like faces and voices were found to significantly promote positive emotional responses to the robots (60). People were also found to report feeling more comfortable around human-like robots and to perceive them as more useful (61). The human-like appearance of an agent even affected the attribution of credit and blame in human-robot teams: automated team-mates who appeared more human-like were relied upon more and attributed more credit for the output than those who didn't (62). Yet, further studies show that the effects of human-like appearances remain unclear. An interesting such counterpoint comes from an experiment which shows that anthropomorphic robots trigger more compensatory consumer responses - yet, not because people liked the human-like robots better, but because they felt discomfort in their presence and perceived them as a threat to human identity (63). And, as was mentioned earlier, participants in another study disliked robots who were described as able to feel (33). More generally, a study on the roots of the behaviors documented by the CASA paradigm failed to find support for the claim that people are more likely to react socially to anthropomorphic characters (22). Unfortunately, a full discussion of the effects of anthropomorphism is beyond the scope of this review. While the selected findings presented here aim to show that appearance seems to matter, we refer to other reviews for a full discussion of the factors involved (see e.g. (64)), as well as of the ethical and societal impacts of anthropomorphic machines (see e.g. (65)).

2.3 The importance of task type

Studies show that humans can sometimes have very strong reactions when a previously human-human interaction is transformed into an interaction with an automated agents. An example of this comes from a field experiment which varied whether sales calls for financial services were made by humans or by chat-bots (66). When the caller's identity was not revealed to the customer, chat-bots made about as many sales as experienced human sales workers. However, if the chat-bot's identity was revealed prior to the call, purchase rates fell by almost 80%. As the follow-up survey revealed, the customers claimed that the chat-bots were less knowledgeable and less empathetic - but only if they knew that they were interacting with chat-bots. However, both the identical sales rates and an analysis of the call data showed no performance differences. Similarly, when participants in another experiment were asked to rate identical creative works allegedly produced by humans or by automated agents, they rated automated works as less morally authentic than the identical human works (67).

Further studies point to a potential reason for these effects: the type of task, and the perceived aptness of using automated agents for such tasks. Indeed, in an experiment investigating the attitudes toward the outsourcing of tasks to robots, people were found to react negatively to the outsourcing of social tasks to robots, but not to the outsourcing of tasks perceived as analytical (68). This distinction was confirmed by several other studies (see e.g. (69; 70; 61)). Additionally, people tend to react to robots more positively if the appearance and demeanor of a robot match the task it is used for (71; 18).

3 Sharing decision authority with automated agents

The increasing involvement of automated agents in decision-making has attracted considerable research interest. The experiments covered in this section all investigate some form of the same general question: Are humans willing to accept the involvement of automated agents in decision-making? And, if so, to what extent? Unlike the research questions, however, the findings of these studies often are not similar. Indeed, while several studies document that humans are averse to delegating decision tasks to automated agents (a phenomenon called *algorithm aversion*), other studies find that they prefer automated advice to human advice (*algorithm appreciation*) or that they over-rely on automated decision-making supports, failing to correct for their mistakes (*automation bias*) or to properly monitor them (*automation-induced complacency*).

In an attempt to reconcile these seemingly inconsistent findings, we propose to categorize the contexts in which they occur according to the distribution of agency within the interaction. That is, we propose to distinguish between situations (a) in which humans make decisions or produce outputs jointly with automated agents (section 3.1), (b) in which the primary or exclusive authority over the decision is delegated to the automated agent (section 3.2), who might then make a decision which affects a human (section 3.5) and (c) in which the human retains the

primary/exclusive authority over the decision (sections 3.3 and 3.4). Fig. 1 illustrates these categories. Next to allowing us to shed a light on potential reasons for inconsistent findings, these categories could be used to formulate concrete testable hypotheses in future research.

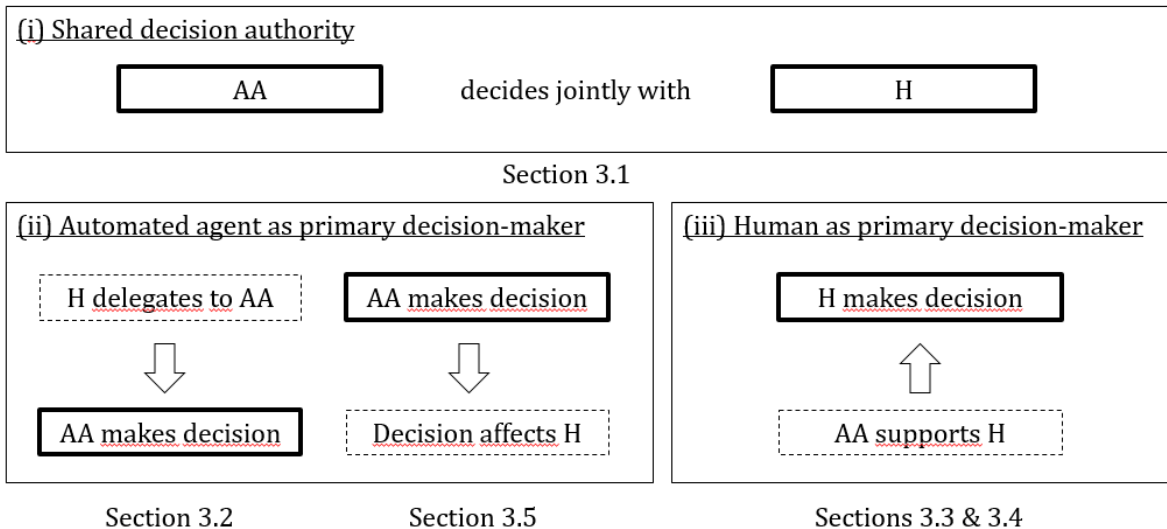


Figure 1: THE INTERACTION OF HUMANS (H) AND AUTOMATED AGENTS (AA) IN DECISION-MAKING. Illustration of different categories of interactions between human and automated agents in decision-making. Categories are separated according to the identity of the primary decision-making authority. The sections of discussion in the text are indicated below.

3.1 The collaboration of humans and automated agents in teams

The growing capacities and the increasing autonomy of automated agents have vastly augmented the roles automated agents can take in the workplace; from mere tools to be used by humans to increasingly autonomous collaborators in their own right. A number of observatory studies closely investigate the introduction of such automated agents into existing organizations (see e.g. 72; 73; 74). Generally, these studies show that the introduction of automated team-mates goes beyond a simple replacement process, and brings with it a number of challenges. Indeed, the introduction of robots in particular into unstructured work environments remains challenging from a technological perspective (75). However, systematic experimental studies of the behavior of humans in hybrid human-machine teams remain rare, though we expect them to grow in number as the technological frontier advances. For the moment, let us deliberate on a few interesting findings to be mentioned here.

First, with regards to the performance of human-machine teams, some studies find that humans extend less effort when their team-mates are automated (76; 56). This might be due to the reduced social response - namely, to the lack of altruism and social pressure humans feel towards

their automated team-mates (56). On the other hand, a different experiment found that participants kept a larger share of the work for themselves when they collaborated with robots rather than human team-mates (77). Similarly, in a field experiment, credit officers extended more effort when they expected an input from a decision-making support system rather than from a human co-worker (78).

Automated coworkers may also affect the distribution of responsibility in a team. Indeed, an experiment shows that participants were more willing to deflect responsibility to automated rather than to human team-mates (76). On the other hand, in a modified dictator game, no significant differences in taking responsibility were detected between human-human and human-computer teams - though participants in human-computer teams did behave slightly more selfishly. (79). Note that by refraining from shifting responsibility towards computers, these participants forewent an apparently effective way to avoid responsibility: as an earlier study showed, observers of accidents tend to attribute less responsibility to a company if technology was involved in the accident (80). Similarly, observers in another study attributed less fault to a doctor who followed the (bad) recommendation of an automated decision aid (47).

Finally, several studies suggest that having an automated team-mate may affect the interaction among the humans in the team (81; 82; 83). For example, one study found that the human team-mates were more likely to engage in social conversations and perceived the group more positively if the robotic teammates expressed vulnerability (e.g. by admitting mistakes) (81). With the important caveat that these studies did not compare if the reactions triggered by robots are different to those that would be triggered by similarly acting humans in the group, these findings suggest that robots might be able to help people collaborate better.

In summary, the existing evidence on the collaboration of humans with relatively autonomous automated agents team settings does not yet allow for the formulation of robust behavioral regularities. Further studies will be needed in this area, in particular on the topics of the impact of automated team-mates on team performance, responsibility attribution and group dynamics.

3.2 Algorithm aversion: The aversion to delegating to automated agents

Let us now turn to the situation where the authority over a decision or the production of an output is ceded, partially or fully, to an automated agent. The question which arises here is whether humans are willing to make this delegation of authority. In an influential series of experiments investigating this question, participants were tasked with a number of forecasting tasks (regarding the future success of MBA graduates and the future development of airline passenger counts), and given the choice between making the predictions themselves or delegating the task to an algorithmic forecaster (84). The participants were aware that the algorithm had made some mistakes in the forecasts - but it still consistently outperformed the human participants. Nevertheless, the participants were more likely to choose to make the prediction themselves. This costly mistake is an example of a behavior the authors of the study termed

algorithm aversion. The preference of humans to rely on themselves rather than to use superior automated decision-making supports had also been documented previously in other contexts, including when driving a car (85) or solving a visual detection task (86).

Importantly, algorithm aversion does not simply imply that people prefer to make decisions themselves. Given the choice between receiving advice from a human or an automated agent, people were more likely to go with the human for joke recommendations (87), medical tasks (88; 89) and when the task was unknown (70). Doctors were perceived as less qualified by observers if they delegated decisions to automated decision support, but not when they sought advice from human colleagues (90; 91).

Algorithm aversion affects decision-making: Experimental participants were more likely to follow medical advice when it came from a human provider rather than from an AI agent (89), and gave more weight to investment advice by humans rather than a statistical forecast (92). In an observatory study (93), managers at a clothing retailer were prone to reject a decision support system's advice regarding markdowns for sales. And, indeed, the debate about when people would do better if they followed algorithmic advice dates back decades (94; 95).

Multiple possible causes of algorithm aversion, or more generally of the "disuse, or the neglect or underutilization of automation" (96), have been proposed by the literature. A first candidate is low trust in automated agents. Indeed, people were found to be particularly unlikely to use automated decision-making aids in uncertain domains (97), where trust is arguably more important. In the study on the use of investment advice, participants stated that they trusted automated agents (slightly) less - though this did not predict whether they actually followed the automated advice (92). A potential reason for the lack of trust in automation is advanced by B. Dietvorst and coauthors (84; 98; 97), who argue that the cause of algorithm aversion is seeing an algorithm err. According to this argument, while people are willing to forgive humans for making mistakes, this leniency is not extended to algorithms. Indeed, earlier studies also show that while participants initially trusted automated agents more than humans and preferred to use them, they lose trust quickly and prefer human aids after seeing the automated agents make mistakes (86). Machine errors have been shown to cause low trust in machines, particularly if error rates are not constant (99). Particularly strong support for the argument of the loss of trust after seeing an algorithm err comes from a more recent study, which finds no support for the hypothesis that human advice is generally preferred to automated advice, but shows that after receiving incorrect advice the utilization of automated advice decreases significantly more than the utilization of human advice (100).

However, findings from other studies cast doubt on these arguments. For instance, the resistance to the use of an AI health provider documented in a series of experiments persisted even when the AI agent was specifically described as being superior to the human regarding the number of complications/accurate diagnoses ((89), study 3c). Neither the (insignificant) differences in trust extended towards automated agents or humans nor the perceived utility of the agent were able to explain the preference for the human decision-maker documented in another experiment

(101). And, two studies of social robots even found that people did either not mind the robots' mistakes (102) or actually preferred robots who made mistakes to those who performed flawlessly (103). Note that, presumably, these last findings were related to the task context: While the above mentioned studies which document salience of machine errors mostly used analytical tasks (e.g., (84; 92)), these robots were used for social tasks. It seems possible that there is some kind of interaction effect between task characteristics and sensitivity to errors: Humans may not be willing to accept automated mistakes in analytical contexts, but they seem to accept or even prefer fallible machines in more social contexts - after all, it would only make the machines appear more human. More generally, the findings of algorithm appreciation and automation bias to be discussed in Sections 3.3 & 3.4 show that people do not by and large distrust machines. Indeed, some papers in these strands of literature document instances where humans trust automated agents too much. Trust might also depend on the anthropomorphic appearance of the automated agent (for an example with autonomous vehicles see (104), for a general overview of the topic of trust in machines see e.g. (105; 106)).

A second potential cause of algorithm aversion is advanced by another experiment (101). Here, participants are given the choice between delegating a calculation task to another human or to an algorithm. The earnings for the task are given to a third person, placing the task in the moral domain. Participants were not only significantly more likely to go with the human decision-maker, but even punished others who chose the algorithm. As neither the trust extended towards the agent (measured with a trust game) nor its perceived utility can explain these findings, the authors argue, people might exhibit a *per se* aversion to the use of automated decision-making agents in the moral domain. Another series of experiments, also documents that humans are averse to automated agents making moral decisions (107). Interestingly, this finding holds irrespective of whether the decisions made are favorable for the third party affected by them. A potential insight into the reasons behind this preference comes from another series of experiments (89): Here, the authors propose that the cause of algorithm aversion is "uniqueness neglect", i.e. the concern that an automated agent is unable to account for an individual's unique characteristics and circumstances. Testing this hypothesis in the healthcare context, the authors find that a participant's perceived sense of uniqueness is able to predict their aversion to the use of medical AI. However, note that the aversion to the use of algorithms in the moral domain does not seem to be a satisfactory explanation for the aversion documented in the studies by B. Dietvorst and coauthors, where algorithms are used to forecast the success of MBA graduates and the development of airline passengers (84), standardized test scores (98) and for a number of deliberately abstract judgement tasks (97) - all tasks outside of the moral domain.

In summary, a number of studies in different contexts and with different participants have documented that humans are averse to delegating tasks to automated agents. This aversion seems to occur in particular where computers are fallible, though there is also evidence that people sometimes don't care about or even value mistakes. Further, the aversion to the use of algorithms seem to be particularly pronounced in moral contexts, and might be caused by the perception that machines are unable to account for human individuality. However, we feel that there is an-

other important factor to note. In almost all the studies we cited above, the participants are found to exhibit algorithm aversion when they are given the choice between a human *or* an automated decision-maker ((84; 101; 88; 87; 70); a notable exception is (92)). A different situation arises when an automated decision-making system is used not to replace the human decision-maker, but to assist him. As we will discuss in the following section, there is evidence which suggests that this factor can not only remedy algorithm aversion, but even trigger a positive preference of automated decision-making support over human aid.

3.3 Algorithm appreciation: The preference for algorithmic advice

Consider the situation where a human remains the primary decision-maker who is supported by an automated agent. Intriguingly, even in the studies which document algorithm aversion, there is evidence that a framing of the agent's role as advisory can remedy algorithm aversion. For example, in (98), people are more likely to use a forecasting algorithm when they are able to intervene and modify its advice (even if just slightly). Participants are less averse to the use of automated agents in medical and other moral decisions when the human decision-makers are presented with a recommendation, rather than a decision (88; 107). An experimental study even documents the full elimination of algorithm aversion when the medical AI is framed as providing support to the human health care provider rather than replacing it ((89), study 9). A vignette study in a consumer and medical setting finds that participants appreciate professionals' use of automated decision aids, but react negatively to a full delegation of decision authority (108).

Some studies show that the change from a situation where the automated agents replaces the human to the one where it supports the human can indeed trigger a preference for automated help over human help. In the leading series of experiments on this phenomenon by Logg et al., participants are asked to make a number of quantitative judgments (visual estimation of a person's weight, prediction of physical attractiveness, prediction of song popularity) under uncertainty, and are offered advice from either humans or automated agents (109). The participants consistently choose to receive and put more weight on automated rather than human advice. The authors term this behavior *algorithm appreciation*. Support for the appreciation of automated advice can also be found in an earlier experiment (110): Here, when participants were provided with (correct) human advice and (incorrect) algorithmic advice for a legal case, they very frequently relied on the algorithmic advice, in particular when it was given in production rule form (replicating a previous finding from (111)). Outside of the lab, people were found to be responsive to algorithmic advice when choosing health care plans in a randomized control trial (however, the study does not compare this to responsiveness to human advice) (112).

Interestingly, even participants in the control condition (i.e., who did not see the model perform) - of the leading experiment that documented algorithm aversion (84) stated that they were more confident in the model's rather than in human forecasts. Similarly, in the study that generally documents the aversion of people to machines making moral decisions (107), in one of the

vignettes participants are asked to choose who they would prefer to make a moral decision: a human doctor, an "autonomous statistics-based computer system" (107, p. 26), or the doctor advised by this system. Participants were most likely to choose the last option. In the words of the authors, "[t]hese results suggest that most people are willing to have machines involved in moral decisions, as long as they are not the ones to make the actual decisions" (107, p. 30).

A possible factor which leads to algorithm aversion or algorithm appreciation is the perception of the relative capabilities of the human and the automated agent, accurate or not. This is illustrated by an interesting replication of the airline passenger forecasting task of (84) in (109). The replication shows that when participants choose between themselves and the algorithm, overconfidence in one's own abilities can lead participants to exhibit algorithm aversion (as in (84)). In contrast, choosing between another human and the algorithm can lead to algorithm appreciation (109). A second difference between the two studies is that in (84) participants have seen the algorithm make mistakes, while participants in (109) received no information on performance prior to making their decision. Finally, the importance of perceived capabilities seems mirrored in the finding that experts are more likely to discard automated advice than are non-experts (see e.g. 109, experiment 4; 113).

In summary, there seem to be decision-making situations in which humans are either indifferent between humans and automated agents or even prefer the automated agents. We hypothesize that this could be due to an important factor distinguishing many of the studies: the retention of at least some form of human determination. So far, the evidence seems to suggest that humans are averse to fully giving up their decision authority, yet appreciative of automated advice when they retain (or feel that they retain) the ultimate authority over the decision. However, this hypothesis rests on relatively few studies as well as some corollary results. Further targeted research will be necessary to properly investigate it.

3.4 Automation bias: The over-reliance on automated support

The consideration of the situation where automated agents are used to provide decision support to humans allows us to discuss another phenomenon documented by the literature: *automation bias*, or the over-reliance on automated decision-making support. In contrast to the previous two sections, in these studies people are not asked to choose between receiving automated or human advice, but are asked to complete a task with the help of automated support. In a number of studies, researchers have repeatedly documented that in such instances, people frequently over-rely on an automated agent's advice, failing to realize when it is wrong and when they should intervene (for literature reviews see e.g. 6; 7). Note that, in fact, this literature describes two related phenomena: automation bias, i.e. the failure to intervene, and *automation-induced complacency*, the failure to appropriately monitor automated support. We will refer to both of these phenomena using the broader term of automation bias (following (6)).

Automation bias occurs when people treat the automated agent "as a heuristic replacement

for vigilant information seeking and processing” (114, p.205). It manifests itself in errors of commission, i.e. misguided actions based on a false alarm by the automated system, or errors of omission, i.e. the missing of critical events when the automated system fails to flag them (115). In one of the earliest studies of this phenomenon, pilots were found to be more easily misled by (faulty) automated checklists than by simple paper checklists (116). This result was then replicated with a flight simulation task in the lab (115). Here, participants in automated conditions were more likely to miss events not flagged by the software (i.e. to commit errors of omission), and more likely to act on incorrect software advice (errors of commission) compared to flying under manual control. Both laypeople and experts are susceptible to automation bias, as was documented in a number of flight simulation studies with experienced pilots (117; 118; 119). Similarly, in a study with experienced air traffic controllers, significantly fewer controllers detected a conflict that the system had failed to flag compared to when conflicts were handled manually (120).

Later studies document the overreliance on automated decision support outside of the domain of aviation. For instance, the use of a spellcheck program was shown to prompt people with high writing skills to miss spelling errors and to falsely change correct spellings (121). In health care, the reliance on incorrect automated advice was shown to lower the decision accuracy of physicians reading EKG charts (122). The sensitivity of professional film readers reading mammograms decreased if they used computer-aided detection systems - causing them to miss cancers if the system had failed to mark them (123). This result was replicated in a follow-on study which in addition found that non-cancerous objects were more likely to be marked as cancers if the system had falsely flagged them as such (124). Further, automation bias and automation-induced complacency have also been documented in process control (125; 126; 127) and command and control situations (128).

In summary, the evidence on automation bias shows that while automated decision-making aids can improve decision quality when they are correct, they may also lower decision quality when they are incorrect, because people are nevertheless prone to rely on them.

Automation bias seems to be more likely to occur in situations of high cognitive load, which may stem from task complexity, multitasking or from time pressure (129). Notably, experience with automated systems or with sharing the load with teammates do not seem to remedy automation bias (130; 131). Indeed, automation bias seems not to stem from general inattention, but from the monitoring of the automated agent having lower priority and receiving less attention than other competing tasks (6). In line with this, an experiment has documented that providing people with variable priority attention training reduces automation bias (132). A further effective way to reduce automation bias seems to be highlighting the responsibility and accountability of the monitoring person (133; 134).

3.5 The acceptance of managerial decisions by automated agents

To end our discussion of the ways in which humans react to automated agents in decision-making, let us consider whether humans are willing to accept automated agents making decisions that concern them. This question is gaining importance as more managerial tasks are performed by automated agents (see e.g. 21; 69)). However, the experimental evidence on whether and when humans are willing to accept automated managerial decisions remains limited - though some interesting findings seem worth mentioning.

First, an obedience experiment demonstrates that people are willing to obey automated managers (135). Here, the manager instructing participants to complete a tedious task is either a human or a robot. While participants protested significantly more frequently, earlier and quit the task sooner when instructed by a robot, they were still willing to obey it. Indeed, about half of the participants in this condition continued with the tedious task for the full duration of the study, even after trying to protest. A noteworthy field experiment with Alibaba warehouse workers shows that humans may sometimes even prefer automated managers to human ones (136). Here workers perceived pick up lists distributed by an algorithm as fairer than those handed out by human managers. As a result warehouse productivity increased in the automated conditions. A lab experiment using a lego set assembly task further documents that people are content to work with a robot allocating tasks within the team (77). Additionally, this study finds that both subjective and objective measures of participants' satisfaction increased with increased autonomy of the automated agent. Workers preferred to have the robot make scheduling decisions, and spent less time on rescheduling tasks the more autonomous the robot was.

However, another experiment in a similar setting finds the opposite results, namely that participants disliked robotic managers (62): the robot managers were more likely to be blamed for mistakes, and relied upon less. Importantly, while in (77) the robots actually performed managerial tasks, in (62) they performed supporting tasks within the team (e.g., carrying the parts), but were exogenously assigned managerial status. Hence, as speculated by the authors in (62), the aversion to robot managers may be due to a mismatch in skills and authority, but not due to the robotic nature of the supervisor itself. The acceptance of automated managerial decisions might also depend on the context of the task. Accordingly, (69) find that human and automated managerial decisions are perceived as equally fair for analytical tasks - but not for social tasks, where the human decisions are perceived as fairer. In addition, the studies discussed in 2.2 found that automated decisions elicit a lower emotional response, which suggests that automated managers might be able to make unpopular decisions with less fear of a backlash. Indeed, there is lab evidence that people perceive coercive decisions labelled as made by a computer as less unfair than identical decisions labelled as made by a human (48).

4 Outline for future research

In summary, the experimental evidence reviewed in this article shows that human-computer interaction is different from human-human interaction. It contains less emotions, and is less affected by social concerns. The answer to the question of when humans are willing to accept automated interaction partners seems less clear: Some experiments find that people prefer automated advice to human advice, (over-)rely on automated advice and are willing to follow automated managers. These findings led to the establishment of the phenomena of algorithm appreciation and automation bias. Others document that humans prefer human decision-makers over equally qualified (or even better) algorithmic support, and hence argue for the phenomenon of algorithm aversion. Hence, the reviewed studies not only show what we know about human-computer interaction, but point out a number of open questions and avenues for future research. In particular, what emerges is a clear need for integrative studies, which simultaneously research the appreciation of and aversion to automated agents to understand the factors triggering one or the other. Such studies would then allow the formulation of a more generalized theoretical framework of how humans interact with automated agents.

For the moment, allow us to formulate a few hypotheses on the potential reasons behind these different findings. First, note that while humans seem willing to accept automated agents in areas considered more objective or analytical, they seem reluctant to do so in areas considered social or moral. One explanation for this could be normative preferences, i.e. the view that some decisions simply should be made by humans (see e.g. 101, who speak of a *per se* aversion to the use of algorithms in the moral domain). A different possible explanation could be distinct performance expectations. (Over-)confidence in one's own abilities or in the general abilities of humans to make complex social or moral decisions could hence form the basis of the aversion to the use of automated agents in such domains (see e.g. 89, who argue for 'uniqueness neglect' as the cause of algorithm aversion). In other words, people might be more willing to accept automated agents in analytical tasks because that is where they expect them to do well. Hence, differential performance expectations could both stand behind the repeated finding that task type matters, and the findings of humans being willing to accept automated agents sometimes yet averse to their use other times.

Another factor which could trigger algorithm aversion in one case, and algorithm appreciation or automation bias in the other is the distribution of agency in the interaction. As we have discussed, there is evidence that people seem particularly appreciative of automated agents when they are framed as providing support rather than as independently making the decision. And, indeed, a number of studies which investigate algorithm aversion find that the aversion decreases or fully disappears when the principal agency is framed to remain with the human. However, studies which purposefully investigate the impact of this factor remain lacking. Hence, for example, it remains an open question whether the degree of human involvement matters - is nominative human involvement enough to avoid algorithm aversion, or does the human need to retain real determinative capacity? And, if nominative human involvement were sufficient to

avoid the bias, would it be ethically and/or legally defensible to mislead people into thinking that they remain involved in decision-making?

5 Discussion and policy insights

The answers to these and similar questions could soon prove to have real-world impact. Recent technological advances have caused extensive discussions of the risks and benefits of automation (see e.g. 137; 138), and have been accompanied by repeated calls for regulation (see e.g. 139; 140; 141). Both public opinion (142) and policy-makers seem to support stricter regulation of automation, in particular with regards to AI (see e.g. 143, for the EU, 144, for China, and 145, for the OECD). The empirical research reviewed in this article can provide a number of insights for the development of such regulation.

First, while the ever-moving technological frontier constantly offers new areas for automation, the research reviewed here shows that the question of where automation is beneficial cannot be answered solely on the basis of technical considerations. The evidence shows that interacting with automated agents reduces the emotional and social response of humans. Hence, where social rules or emotions form obstacles, automation can bring benefits beyond those of time and labour savings - for example, by increasing the willingness people to disclose sensitive information like intimate partner violence (51; 52). In contrast, in other situations social rules and expectations are desirable, and automation can be detrimental. For example, studies on charitable giving have repeatedly shown that social image concerns and social pressure are some of important drivers of charitable giving (146; 147). This suggests that automation might prove harmful here.

A second insight concerns the optimal degree of automation. Numerous studies have documented that fully automated decisions can be more accurate and less discriminatory than human decisions (see e.g. 93; 148; 149; 150; 151). Hence, one might argue, it would be beneficial to take humans off the decision loop, thereby simplifying the decision problem to the task of creating the most accurate and least biased algorithm (see e.g. 152). However, the studies reviewed here show that humans are particularly likely to exhibit algorithm aversion when they are - or when they feel that they are - replaced by automated agents. Entirely removing humans from the decision loop could thus make them particularly likely to mistrust and make inefficiently little use of algorithms. This empirical finding seems to be reflected in official government positions (see e.g. 153) and even in regulation (see e.g. art. 22(1) EU GDPR). However, further studies reviewed in this article also point to a potential downside of retaining human involvement in automated decision-making: people may over-rely on the recommendations produced by automated support systems, and fail to correct for the system's mistakes. For the moment, the reviewed evidence shows that the degree of human involvement constitutes an important factor which affects people's reaction towards automation. Further studies investigating this phenomenon will be necessary to solve the problem of the optimal degree of automation.

A particularly promising avenue of research in this regard seems to be the investigation of the impact of (real or perceived) accountability. Indeed, studies suggest that highlighting the responsibility of a human agent for the outcome of a decision could both decrease algorithm aversion by way of ensuring human intervention (see e.g. 98) and decrease automation bias by ensuring more effective monitoring (see e.g. 133). However, ensuring human accountability has recently become more difficult with the growing use of opaque, 'black-box' algorithms (see e.g. 137; 154)). The obvious solution to this problem seems to lie in increased algorithmic transparency. Behavioral research can again provide valuable insights here. Indeed, experiments show that transparency can also backfire; too much transparency can discourage people from using technology (155; 156), or create incentives to game the system (157). In addition, transparency can be difficult to establish and have unforeseen effects. To illustrate this point, consider a particularly interesting finding from a computer science user study which develops an explanation method for a machine learning algorithm used to distinguish pictures of wolves and huskies (158). To test the explanations, the authors first showed a group of students trained in machine learning a number of classifications the algorithm had made. When the students were shown the raw data (the pictures) as well as the labels the algorithm had proposed, most of them stated that they trusted the algorithm to perform well, even though they could see that it had made some mistakes. Only if, in addition, the authors showed the students an explanation of how the algorithm made these classifications (namely that the snow in the background of the picture was a relevant factor) did the students lose trust in the algorithm - even though they had already known its accuracy rate before receiving the explanation. Hence, explanations matter: simply providing information about the accuracy of algorithm is not sufficient to enable observers to accurately judge an algorithm's validity. Indeed, a recent experiment shows that the way an algorithm is explained seems to affect how people perceive it (159).

More generally, transparency about automated decision-making can be misleading if there is no appropriate comparison to human decision-making. An example of this comes from a study which compares a CV screening algorithm and human HR managers (160). First, the author establishes that the algorithm gives a negative weight to candidates from non-elite universities. Thus, one might conclude, the algorithm is biased and harmful. However, it turned out that the candidates from non-elite schools actually disproportionately benefited from being screened by the algorithm, as the human evaluators would have assessed their credentials even more negatively. A similar finding comes from another study which finds that an automated recidivism risk score is biased, but not that the involvement of human decision-makers would result in any less biased decisions (161). On top of the facts of how unbiased decisions are, perceptions of those decisions by people are important. People may perceive algorithmic decisions as less than fair even when they objectively are (162) or think that automated decisions are fairer than identical human decisions (136). Further research will be needed to fully establish the impacts of transparency on automated decision-making.

In summary, the behavioral research reviewed in this article provides a number of important insights into how humans interact with automated agents, and can serve to inform political

discussion about how best to regulate automation. With the evidence gathered here, and by highlighting potential inconsistencies and avenues for future research, we hope to stimulate further research on this topic.

References

- [1] Rahwan, I. *et al.* Machine behaviour. *Nature* **568**, 477–486 (2019).
- [2] Whittmore, R. & Knafl, K. The integrative review: updated methodology. *Journal of advanced nursing* **52**, 546–553 (2005).
- [3] Thompson, S. G. & Pocock, S. J. Can meta-analyses be trusted? *The Lancet* **338**, 1127–1130 (1991).
- [4] Jussupow, E., Benbasat, I. & Heinzl, A. Why are we averse towards algorithms? a comprehensive literature review on algorithm aversion. In *ECIS* (2020).
- [5] Burton, J. W., Stein, M.-K. & Jensen, T. B. A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making* **33**, 220–239 (2020).
- [6] Parasuraman, R. & Manzey, D. H. Complacency and bias in human use of automation: an attentional integration. *Human Factors* **52**, 381–410 (2010).
- [7] Alberdi, E., Strigini, L., Povyakalo, A. A. & Ayton, P. Why are people's decisions sometimes worse with computer support? In *International Conference on Computer Safety, Reliability, and Security*, 18–31 (Springer, 2009).
- [8] March, C. The behavioral economics of artificial intelligence: Lessons from experiments with computer players. *CESifo Working Paper No. 7926*. (2019). Available at SSRN: <https://ssrn.com/abstract=3485475>.
- [9] Bruner, J. The narrative construction of reality. *Critical Inquiry* **18**, 1–21 (1991).
- [10] Heider, F. & Simmel, M. An experimental study of apparent behavior. *The American journal of psychology* **57**, 243–259 (1944).
- [11] Nass, C. & Moon, Y. Machines and mindlessness: social responses to computers. *Journal of social issues* **56**, 81–103 (2000).
- [12] Nass, C., Green, N. & Moon, Y. Are machines gender neutral? gender-stereotypic responses to computers with voices. *Journal of Applied Social Psychology* **27**, 864–876 (1997).
- [13] Nass, C., Fogg, B. J. & Moon, Y. Can computers be teammates? *International Journal of Human-Computer Studies* **45**, 669–678 (1996).
- [14] De Laere, K. H., Lundgren, D. C. & Howe, S. R. The electronic mirror: human-computer interaction and change in self-appraisals. *Computers in Human Behavior* **14**, 43–59 (1998).

- [15] Fogg, B. J. & Nass, C. How users reciprocate to computers: an experiment that demonstrates behavior change. In *CHI'97 Extended Abstracts on Human Factors in Computing Systems*, 331–332 (1997).
- [16] Katagiri, Y., Nass, C. & Takeuchi, Y. Cross-cultural studies of the computers are social actors paradigm: The case of reciprocity 1558–1562 (2001).
- [17] Tay, B. T. C., Park, T., Jung, Y., Tan, Y. K. & Wong, A. H. Y. When stereotypes meet robots: the effect of gender stereotypes on people’s acceptance of a security robot. In *International Conference on Engineering Psychology and Cognitive Ergonomics*, 261–270 (Springer, 2013).
- [18] Eyssel, F. & Hegel, F. (s)he’s got the look: Gender stereotyping of robots. *Journal of Applied Social Psychology* **42**, 2213–2230 (2012).
- [19] Tay, B., Jung, Y. & Park, T. When stereotypes meet robots: the double-edge sword of robot gender and personality in human–robot interaction. *Computers in Human Behavior* **38**, 75–84 (2014).
- [20] Bartneck, C. *et al.* Robots and racism. In *Proceedings of the 2018 ACM/IEEE international conference on human-robot interaction*, 196–204 (2018).
- [21] Aharoni, E. & Fridlund, A. J. Social reactions toward people vs. computers: how mere labiles shape interactions. *Computers in Human Behavior* **23**, 2175–2189 (2007).
- [22] Lee, E.-J. What triggers social responses to flattering computers? experimental tests of anthropomorphism and mindlessness explanations. *Communication Research* **37**, 191–214 (2010).
- [23] Salomons, N., van der Linden, M., Strohkorb Sebo, S. & Scassellati, B. Humans conform to robots: Disambiguating trust, truth, and conformity. In *Proceedings of the 2018 acm/ieee international conference on human-robot interaction*, 187–195 (2018).
- [24] Von der Puetten, A. M., Krämer, N. C., Gratch, J. & Kang, S.-H. “it doesn’t matter what you are!” explaining social effects of agents and avatars. *Computers in Human Behavior* (2010).
- [25] Krach, S. *et al.* Can machines think? interaction and perspective taking with robots investigated via fmri. *PloS one* **3** (2008).
- [26] Chaminade, T. *et al.* How do we think machines think? an fmri study of alleged competition with an artificial intelligence. *Frontiers in Human Neuroscience* **6**, 103 (2012).
- [27] Sanfey, A. G., Rilling, J. K., Aronson, J. A., Nystrom, L. E. & Cohen, J. D. The neural basis of economic decision – making in the ultimatum game. *Science* **300**, 1755–1758 (2003).

- [28] McCabe, K., Houser, D., Ryan, L., Smith, V. & Trouard, T. A functional imaging study of cooperation in two-person reciprocal exchange. *Proceedings of the National Academy of Sciences* **98**, 11832–11835 (2001).
- [29] Frith, C. D. & Frith, U. The neural basis of mentalizing. *Neuron* **50**, 531–534 (2006).
- [30] Coricelli, G. & Nagel, R. Neural correlates of depth of strategic reasoning in medial prefrontal cortex. *Proceedings of the National Academy of Sciences* **106**, 9163–9168 (2009).
- [31] Farjam, M. & Kirchkamp, O. Bubbles in hybrid markets: how expectations about algorithmic trading affect human trading. *Journal of Economic Behavior & Organization* **146**, 248–269 (2018).
- [32] Gray, H. M., Gray, K. & Wegner, D. M. Dimensions of mind perception. *science* **315**, 619 (2007).
- [33] Gray, K. & Wegner, D. M. Feeling robots and human zombies: mind perception and the uncanny valley. *Cognition* **125**, 125–130 (2012).
- [34] Rosenthal-von der Pütten, A. M., Krämer, N. C., Hoffmann, L., Sobieraj, S. & Eimler, S. C. An experimental study on emotional reactions towards a robot. *International Journal of Social Robotics* **5**, 17–34 (2013).
- [35] Bartneck, C., Rosalia, C., Menges, R. & Deckers, I. Robot abuse – a limitation of the media equation (2005).
- [36] Slater, M. *et al.* A virtual reprise of the stanley milgram obedience experiments. *PloS one* **1**, e39 (2006).
- [37] Milgram, S. Behavioral study of obedience. *The Journal of Abnormal and Social psychology* **67**, 371 (1963).
- [38] Briggs, G. & Scheutz, M. How robots can affect human behavior: Investigating the effects of robotic displays of protest and distress. *International Journal of Social Robotics* **6**, 343–355 (2014).
- [39] Schniter, E., Shields, T. W. & Sznycer, D. Trust in humans and robots: economically similar but emotionally different. *Journal of Economic Psychology* 102253 (2020).
- [40] Lim, S. & Reeves, B. Computer agents versus avatars: responses to interactive game characters controlled by a computer or other player. *International Journal of Human – Computer Studies* **68**, 57–68 (2010).
- [41] Melo, C. D., Marsella, S. & Gratch, J. People do not feel guilty about exploiting machines. *ACM Transactions on Computer – Human Interaction (TOCHI)* **23**, 1–17 (2016).

- [42] Mandryk, R. L., Inkpen, K. M. & Calvert, T. W. Using psychophysiological techniques to measure user experience with entertainment technologies. *Behaviour & information technology* **25**, 141–158 (2006).
- [43] van't Wout, M., Kahn, R. S., Sanfey, A. G. & Aleman, A. Affective state and decision-making in the ultimatum game. *Experimental brain research* **169**, 564–568 (2006).
- [44] Teubner, T., Adam, M. & Riordan, R. The impact of computerized agents on immediate emotions, overall arousal and bidding behavior in electronic auctions. *Journal of the Association for Information Systems* **16**, 838 (2015).
- [45] Adam, M. T. P., Krämer, J. & Müller, M. B. Auction fever! how time pressure and social competition affect bidders' arousal and bids in retail auctions. *Journal of Retailing* **91**, 468–485 (2015).
- [46] Leyer, M. & Schneider, S. Me, you or ai? how do we feel about delegation. In *Proceedings of the 27th European Conference on Information Systems (ECIS)* (Springer, 2019). URL https://aisel.aisnet.org/ecis2019_rp/36.
- [47] Pezzo, M. V. & Pezzo, S. P. Physician evaluation after medical errors: does having a computer decision aid help or hurt in hindsight? *Medical Decision Making* **26**, 48–56 (2006).
- [48] Shank, D. B. Perceived justice and reactions to coercive computers. In *Sociological Forum*, vol. 27, 372–391 (Wiley Online Library, 2012).
- [49] Adam, M. T. P., Teubner, T. & Gimpel, H. No rage against the machine: how computer agents mitigate human emotional processes in electronic negotiations. *Group Decision and Negotiation* **27**, 543–571 (2018).
- [50] Lucas, G. M., Gratch, J., King, A. & Morency, L.-P. It's only a computer: virtual humans increase willingness to disclose. *Computers in Human Behavior* **37**, 94–100 (2014).
- [51] Ahmad, F. *et al.* Computer-assisted screening for intimate partner violence and control: a randomized trial. *Annals of Internal Medicine* **151**, 93–102 (2009).
- [52] Humphreys, J., Tsoh, J. Y., Kohn, M. A. & Gerbert, B. Increasing discussions of intimate partner violence in prenatal care using video doctor plus provider cueing: a randomized, controlled trial. *Women's health issues* **21**, 136–144 (2011).
- [53] Moon, Y. & Nass, C. Are computers scapegoats? attributions of responsibility in human – computer interaction. *International Journal of Human-Computer Studies* **49**, 79–94 (1998).

- [54] Moon, Y. Don't blame the computer: when self-disclosure moderates the self-serving bias. *Journal of Consumer Psychology* **13**, 125–137 (2003).
- [55] Cohn, A., Gesche, T. & Maréchal, M. A. Honesty in the digital age (2018).
- [56] Corgnet, B., Hernán-Gonzalez, R. & Mateo, R. Rac(g)e against the machine?: Social incentives when humans meet robots. *Social Incentives When Humans Meet Robots (January 28, 2019)*. GATE WP (2019).
- [57] Gratch, J., Wang, N., Gerten, J., Fast, E. & Duffy, R. Creating rapport with virtual agents. In *International Workshop on Intelligent Virtual Agents*, 125–138 (Springer, 2007).
- [58] Bickmore, T., Gruber, A. & Picard, R. Establishing the computer – patient working alliance in automated health behavior change interventions. *Patient Education and Counseling* **59**, 21–30 (2005).
- [59] Short, E., Hart, J., Vu, M. & Scassellati, B. No fair!! an interaction with a cheating robot. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 219–226 (IEEE, 2010).
- [60] Zhang, T. *et al.* Service robot feature design effects on user perceptions and emotional responses. *Intelligent service robotics* **3**, 73–88 (2010).
- [61] Castelo, N. *Blurring the Line Between Human and Machine: Marketing Artificial Intelligence*. Ph.D. thesis, Columbia University (2019).
- [62] Hinds, P. J., Roberts, T. L. & Jones, H. Whose job is it anyway? a study of human-robot interaction in a collaborative task. *Human-Computer Interaction* **19**, 151–181 (2004).
- [63] Mende, M., Scott, M. L., van Doorn, J., Grewal, D. & Shanks, I. Service robots rising: how humanoid robots influence service experiences and elicit compensatory consumer responses. *Journal of Marketing Research* **56**, 535–556 (2019).
- [64] Złotowski, J., Proudfoot, D., Yogeewaran, K. & Bartneck, C. Anthropomorphism: opportunities and challenges in human–robot interaction. *International journal of social robotics* **7**, 347–360 (2015).
- [65] Darling, K. 'who's johnny?' anthropomorphic framing in human-robot interaction, integration, and policy. *Anthropomorphic Framing in Human-Robot Interaction, Integration, and Policy (March 23, 2015)*. *ROBOT ETHICS* **2** (2015).
- [66] Luo, X., Tong, S., Fang, Z. & Qu, Z. Frontiers: machines vs. humans: The impact of artificial intelligence chatbot disclosure on customer purchases. *Marketing Science* **38**, 937–947 (2019).

- [67] Jago, A. S. Algorithms and authenticity. *Academy of Management Discoveries* **5**, 38–56 (2019).
- [68] Waytz, A. & Norton, M. I. Botsourcing and outsourcing: robot, british, chinese, and german workers are for thinking — not feeling — jobs. *Emotion* **14**, 434 (2014).
- [69] Lee, M. K. Understanding perception of algorithmic decisions: fairness, trust, and emotion in response to algorithmic management. *Big Data & Society* **5**, 2053951718756684 (2018).
- [70] Hertz, N. & Wiese, E. Good advice is beyond all price, but what if it comes from a machine? *Journal of Experimental Psychology: Applied* (2019).
- [71] Goetz, J., Kiesler, S. & Powers, A. Matching robot appearance and behavior to tasks to improve human-robot cooperation. In *The 12th IEEE International Workshop on Robot and Human Interactive Communication, 2003. Proceedings. ROMAN 2003.*, 55–60 (Ieee, 2003).
- [72] van den Broek, E. Hiring algorithms: an ethnography of fairness in practice. In *The Future of Work* (ICIS, 2019). ICIS 2019 Proceedings No.2177.
- [73] Lebovitz, S., Levina, N. & Lifshitz-Assaf, H. Doubting the diagnosis: how artificial intelligence increases ambiguity during professional decision making (2019). Available at SSRN 3480593.
- [74] Stubbs, K., Hinds, P. J. & Wettergreen, D. Autonomy and common ground in human-robot interaction: a field study. *IEEE Intelligent Systems* **22**, 42–50 (2007).
- [75] Haddadin, S. & Croft, E. Physical human – robot interaction. In *Springer Handbook of Robotics*, 1835–1874 (Springer, 2016).
- [76] Domeinski, J., Wagner, R., Schöbel, M. & Manzey, D. Human redundancy in automation monitoring: effects of social loafing and social compensation. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 51, 587–591 (SAGE Publications Sage CA: Los Angeles, CA, 2007).
- [77] Gombolay, M. C., Gutierrez, R. A., Clarke, S. G., Sturla, G. F. & Shah, J. A. Decision-making authority, team efficiency and human worker satisfaction in mixed human–robot teams. *Autonomous Robots* **39**, 293–312 (2015).
- [78] Paravisini, D. & Schoar, A. The incentive effect of scores: randomized evidence from credit committees. Tech. Rep., National Bureau of Economic Research (2013).
- [79] Kirchkamp, O. & Strobel, C. Sharing responsibility with a machine. *Journal of Behavioral and Experimental Economics* **80**, 25–33 (2019).

- [80] Kurtzberg, T. R. & Naquin, C. E. Human reactions to technological failure: How accidents rooted in technology vs. human error influence judgments of organizational accountability. *Organizational Behavior and Human Decision Processes* **93**, 129–141 (2004).
- [81] Traeger, M. L., Sebo, S. S., Jung, M., Scassellati, B. & Christakis, N. A. Vulnerable robots positively shape human conversational dynamics in a human–robot team. *Proceedings of the National Academy of Sciences* **117**, 6370–6375 (2020).
- [82] Short, E. & Mataric, M. J. Robot moderation of a collaborative game: Towards socially assistive robotics in group interactions. In *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 385–390 (IEEE, 2017).
- [83] Strohkorb, S. *et al.* Improving human-human collaboration between children with a social robot. In *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 551–556 (IEEE, 2016).
- [84] Dietvorst, B. J., Simmons, J. P. & Massey, C. Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* **144**, 114 (2015).
- [85] Kantowitz, B. H., Hanowski, R. J. & Kantowitz, S. C. Driver acceptance of unreliable traffic information in familiar and unfamiliar settings. *Human Factors* **39**, 164–176 (1997).
- [86] Dzindolet, M. T., Pierce, L. G., Beck, H. P. & Dawe, L. A. The perceived utility of human and automated aids in a visual detection task. *Human Factors* **44**, 79–94 (2002).
- [87] Yeomans, M., Shah, A., Mullainathan, S. & Kleinberg, J. Making sense of recommendations. *Journal of Behavioral Decision Making* **32**, 403–414 (2019).
- [88] Promberger, M. & Baron, J. Do patients trust computers? *Journal of Behavioral Decision Making* **19**, 455–468 (2006).
- [89] Longoni, C., Bonezzi, A. & Morewedge, C. K. Resistance to medical artificial intelligence. *Journal of Consumer Research* **46**, 629–650 (2019).
- [90] Arkes, H. R., Shaffer, V. A. & Medow, M. A. Patients derogate physicians who use a computer-assisted diagnostic aid. *Medical Decision Making* **27**, 189–202 (2007).
- [91] Shaffer, V. A., Probst, C. A., Merkle, E. C., Arkes, H. R. & Medow, M. A. Why do patients derogate physicians who use a computer-based diagnostic support system? *Medical Decision Making* **33**, 108–118 (2013).

- [92] Önköl, D., Goodwin, P., Thomson, M., Gönül, S. & Pollock, A. The relative influence of advice from human experts and statistical methods on forecast adjustments. *Journal of Behavioral Decision Making* **22**, 390–409 (2009).
- [93] Saez de Tejada Cuenca, A. *Essays on Social and Behavioral Aspects of Apparel Supply Chains* (University of California, Los Angeles, 2019).
- [94] Meehl, P. E. When shall we use our heads instead of the formula? *Journal of Counseling Psychology* **4**, 268 (1957).
- [95] Meehl, P. E. Clinical versus statistical prediction: a theoretical analysis and a review of the evidence. (1954).
- [96] Parasuraman, R. & Riley, V. Humans and automation: use, misuse, disuse, abuse. *Human Factors* **39**, 230–253 (1997).
- [97] Dietvorst, B. & Bharti, S. People reject algorithms in uncertain decision domains because they have diminishing sensitivity to forecasting error (2019). Available at SSRN: <https://ssrn.com/abstract=3424158>.
- [98] Dietvorst, B. J., Simmons, J. P. & Massey, C. Overcoming algorithm aversion: people will use imperfect algorithms if they can (even slightly) modify them. *Management Science* **64**, 1155–1170 (2018).
- [99] Muir, B. M. & Moray, N. Trust in automation. part ii. experimental studies of trust and human intervention in a process control simulation. *Ergonomics* **39**, 429–460 (1996).
- [100] Prael, A. & Van Swol, L. Understanding algorithm aversion: when is advice from automation discounted? *Journal of Forecasting* **36**, 691–702 (2017).
- [101] Gogoll, J. & Uhl, M. Rage against the machine: automation in the moral domain. *Journal of Behavioral and Experimental Economics* **74**, 97–103 (2018).
- [102] Salem, M., Lakatos, G., Amirabdollahian, F. & Dautenhahn, K. Would you trust a (faulty) robot? effects of error, task type and personality on human-robot cooperation and trust. In *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 1–8 (IEEE, 2015).
- [103] Mirnig, N. *et al.* To err is robot: How humans assess and act toward an erroneous social robot. *Frontiers in Robotics and AI* **4**, 21 (2017).
- [104] Waytz, A., Heafner, J. & Epley, N. The mind in the machine: anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology* **52**, 113–117 (2014).

- [105] Lee, J. D. & See, K. A. Trust in automation: designing for appropriate reliance. *Human Factors* **46**, 50–80 (2004).
- [106] Glikson, E. & Woolley, A. W. Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals* (2020).
- [107] Bigman, Y. E. & Gray, K. People are averse to machines making moral decisions. *Cognition* **181**, 21–34 (2018).
- [108] Palmeira, M. & Spassova, G. Consumer reactions to professionals who use decision aids. *European Journal of Marketing* (2015).
- [109] Logg, J. M., Minson, J. A. & Moore, D. A. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes* **151**, 90–103 (2019).
- [110] Dijkstra, J. J. User agreement with incorrect expert system advice. *Behaviour & Information Technology* **18**, 399–411 (1999).
- [111] Dijkstra, J. J., Liebrand, W. B. & Timminga, E. Persuasiveness of expert systems. *Behaviour & Information Technology* **17**, 155–163 (1998).
- [112] Bundorf, K., Polyakova, M. & Tai-Seale, M. How do humans interact with algorithms? experimental evidence from health insurance. Tech. Rep., National Bureau of Economic Research (2019).
- [113] Tazelaar, F. & Snijders, C. The myth of purchasing professionals’ expertise. more evidence on whether computers can make better procurement decisions. *Journal of Purchasing and Supply Management* **10**, 211–222 (2004).
- [114] Mosier, K. L. & Skitka, L. J. Human decision makers and automated decision aids: made for each other? In Parasuraman, R. & Mouloua, M. (eds.) *Automation and Human Performance: Theory and Application*, 201–220 (Routledge, 1996).
- [115] Skitka, L. J., Mosier, K. L. & Burdick, M. Does automation bias decision-making? *International Journal of Human-Computer Studies* **51**, 991–1006 (1999).
- [116] Mosier, K. L., Palmer, E. A. & Degani, A. Electronic checklists: implications for decision making. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 36, 7–11 (SAGE Publications Sage CA: Los Angeles, CA, 1992).
- [117] Mosier, K. L., Skitka, L. J., Heers, S. & Burdick, M. Automation bias: decision making and performance in high-tech cockpits. *The International Journal of Aviation Psychology* **8**, 47–63 (1998).

- [118] Sarter, N. B. & Schroeder, B. Supporting decision making and action selection under time pressure and uncertainty: the case of in-flight icing. *Human Factors* **43**, 573–583 (2001).
- [119] Galster, S. M., Duley, J. A., Masalonis, A. J. & Parasuraman, R. Air traffic controller performance and workload under mature free flight: Conflict detection and resolution of aircraft self-separation. *The international journal of aviation psychology* **11**, 71–93 (2001).
- [120] Metzger, U. & Parasuraman, R. Automation in future air traffic management: effects of decision aid reliability on controller performance and mental workload. *Human Factors* **47**, 35–49 (2005).
- [121] Galletta, D. F., Durcikova, A., Everard, A. & Jones, B. M. Does spell-checking software need a warning label? *Communications of the ACM* **48**, 82–86 (2005).
- [122] Tsai, T. L., Fridsma, D. B. & Gatti, G. Computer decision support as a source of interpretation error: the case of electrocardiograms. *Journal of the American Medical Informatics Association* **10**, 478–483 (2003).
- [123] Alberdi, E., Povyakalo, A., Strigini, L. & Ayton, P. Effects of incorrect computer-aided detection (cad) output on human decision-making in mammography. *Academic Radiology* **11**, 909–918 (2004).
- [124] Alberdi, E., Povyakalo, A. A., Strigini, L., Ayton, P. & Given-Wilson, R. Cad in mammography: lesion-level versus case-level analysis of the effects of prompts on human decisions. *International Journal of Computer Assisted Radiology and Surgery* **3**, 115–122 (2008).
- [125] Bahner, J. E., Elepfandt, M. F. & Manzey, D. Misuse of diagnostic aids in process control: the effects of automation misses on complacency and automation bias. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 52, 1330–1334 (SAGE Publications Sage CA: Los Angeles, CA, 2008).
- [126] Manzey, D., Reichenbach, J. & Onnasch, L. Human performance consequences of automated decision aids: the impact of degree of automation and system experience. *Journal of Cognitive Engineering and Decision Making* **6**, 57–87 (2012).
- [127] Wickens, C. D., Clegg, B. A., Vieane, A. Z. & Sebok, A. L. Complacency and automation bias in the use of imperfect automation. *Human Factors* **57**, 728–739 (2015).
- [128] Rovira, E., McGarry, K. & Parasuraman, R. Effects of imperfect automation on decision making in a simulated command and control task. *Human Factors* **49**, 76–87 (2007).

- [129] Parasuraman, R., Molloy, R. & Singh, I. L. Performance consequences of automation-induced 'complacency'. *The International Journal of Aviation Psychology* **3**, 1–23 (1993).
- [130] Skitka, L. J., Mosier, K. L., Burdick, M. & Rosenblatt, B. Automation bias and errors: are crews better than individuals? *The International Journal of Aviation Psychology* **10**, 85–97 (2000).
- [131] Mosier, K. L., Skitka, L. J., Dunbar, M. & McDonnell, L. Aircrews and automation bias: the advantages of teamwork? *The International Journal of Aviation Psychology* **11**, 1–14 (2001).
- [132] Metzger, U., Duley, J. A., Abbas, R. & Parasuraman, R. Effects of variable-priority training on automation-related complacency: performance and eye movements. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 44, 346–349 (SAGE Publications Sage CA: Los Angeles, CA, 2000).
- [133] Goddard, K., Roudsari, A. & Wyatt, J. C. Automation bias: a systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association* **19**, 121–127 (2012).
- [134] Skitka, L. J., Mosier, K. & Burdick, M. D. Accountability and automation bias. *International Journal of Human-Computer Studies* **52**, 701–717 (2000).
- [135] Cormier, D., Newman, G., Nakane, M., Young, J. E. & Durocher, S. Would you do as a robot commands? an obedience study for human-robot interaction. In *International Conference on Human-Agent Interaction* (2013).
- [136] Bai, B., Dai, H., Zhang, D., Zhang, F. & Hu, H. The impacts of algorithmic work assignment on fairness perceptions and productivity: evidence from field experiments (2020). Available at SSRN: <https://ssrn.com/abstract=355088>.
- [137] Pasquale, F. *The Black Box Society* (Harvard University Press, 2015).
- [138] Cowgill, B. & Tucker, C. E. Economics, fairness and algorithmic bias (2019). Available at SSRN: <https://ssrn.com/abstract=3361280>.
- [139] Citron, D. K. & Pasquale, F. The scored society: due process for automated predictions. *Washington Law Review* **89**, 1 (2014).
- [140] Doshi-Velez, F. *et al.* Accountability of ai under the law: the role of explanation (2017). Preprint at arXiv:1711.01134.
- [141] Wachter, S. & Mittelstadt, B. A right to reasonable inferences: re-thinking data protection law in the age of big data and ai. *Columbia Business Law Review* 494 (2019).

- [142] Grzymek, V. & Puntschuh, M. What Europe knows and thinks about algorithms results of a representative survey. In *Bertelsmann Stiftung Eupinions February 2019* (2019).
- [143] Wallace, N. Europe plans to strictly regulate high-risk AI technology (2020). <https://www.sciencemag.org/news/2020/02/europe-plans-strictly-regulate-high-risk-ai-technology>.
- [144] Roberts, H. *et al.* The Chinese approach to artificial intelligence: an analysis of policy and regulation (2019). Available at SSRN 3469783.
- [145] OECD. Recommendation of the Council on Artificial Intelligence (2019).
- [146] Della Vigna, S., List, J. A. & Malmendier, U. Testing for altruism and social pressure in charitable giving. *The Quarterly Journal of Economics* **127**, 1–56 (2012).
- [147] Andreoni, J., Rao, J. M. & Trachtman, H. Avoiding the ask: A field experiment on altruism, empathy, and charitable giving. *Journal of Political Economy* **125**, 625–653 (2017).
- [148] Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J. & Mullainathan, S. Human decisions and machine predictions. *The Quarterly Journal of Economics* **133**, 237–293 (2018).
- [149] Stevenson, M. Assessing risk assessment in action. *Minnesota Law Review* **103**, 303 (2018).
- [150] Hoffman, M., Kahn, L. B. & Li, D. Discretion in hiring. *The Quarterly Journal of Economics* **133**, 765–800 (2018).
- [151] Gates, S. W., Perry, V. G. & Zorn, P. M. Automated underwriting in mortgage lending: good news for the underserved? *Housing Policy Debate* **13**, 369–391 (2002).
- [152] Miller, A. P. Want less – biased decisions? Use algorithms. *Harvard Business Review* **26** (2018).
- [153] Hancock, M. Artificial intelligence: opportunities and implications for the future of decision making. *Governments Office for Science* (2015).
- [154] Lipton, Z. C. The mythos of model interpretability. *Queue* **16**, 31–57 (2018).
- [155] Kizilcec, R. F. How much information? Effects of transparency on trust in an algorithmic interface. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2390–2395 (2016).

- [156] Lee, M. K., Jain, A., Cha, H. J., Ojha, S. & Kusbit, D. Procedural justice in algorithmic fairness: leveraging transparency and outcome control for fair algorithmic mediation. *Proceedings of the ACM on Human-Computer Interaction* **3**, 1–26 (2019).
- [157] Ederer, F., Holden, R. & Meyer, M. Gaming and strategic opacity in incentive provision. *The RAND Journal of Economics* **49**, 819–854 (2018).
- [158] Ribeiro, M. T., Singh, S. & Guestrin, C. ” why should i trust you?” explaining the predictions of any classifier. In *Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining*, 1135–1144 (2016).
- [159] Cowgill, B., Dell’Acqua, F. & Matz, S. The managerial effects of algorithmic fairness activism. In *AEA Papers and Proceedings*, vol. 110, 85–90 (2020).
- [160] Cowgill, B. Bias and productivity in humans and algorithms: Theory and evidence from resume screening (2018).
- [161] Tan, S., Adebayo, J., Inkpen, K. & Kamar, E. Investigating human+ machine complementarity for recidivism predictions (2018). Preprint at arXiv:1808.09123.
- [162] Lee, M. K. & Baykal, S. Algorithmic mediation in group decisions: Fairness perceptions of algorithmically mediated vs. discussion-based social division. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 1035–1048 (2017).

Appendix: Methodological overview of reviewed studies

The below tables provide details on the studies reviewed in this article. Table 1 gives short methodological notes on the experimental studies. Table 2 briefly describes the mentioned observational studies. Table 3 briefly describes the mentioned literature reviews.

The following should be noted about the classifications used in table 1: Where a study contains multiple experiments, experiments are enumerated following the authors' system. Wherever authors do not explicitly give information about a methodological detail about their study, the table states "N/A". In contrast, the use of "None", means that the authors state that a method was not used. E.g., regarding incentives, "None" signifies that no incentives were used, whereas "N/A" signifies that the paper does not specify either way. Incentives are differentiated into contingent or non-contingent incentives, without further elaboration of whether they consist of monetary payments or non-monetary rewards (such as e.g. course credits). Further, regarding the elicitation measures, the observation of participants' actions and choices is classified as revealed preferences independently of whether they are incentivized or not.

Table 1: EXPERIMENTAL STUDIES

Author	Title	Type	N	Subjects	Incentives	Elicitation measures
Adam, Krämer, et al., 2015	"Auction fever! How time pressure and social competition affect bidders' arousal and bids in retail auctions"	1: Lab	240	Students	Contingent	Revealed preferences & physiological measures
		2: Lab	216	Students	Contingent	Stated, revealed preferences
Adam, Teubner, et al., 2018	"No rage against the machine: how computer agents mitigate human emotional processes in electronic negotiations"	Lab	216	Students	Contingent	Revealed preferences & physiological measures

Table 1: EXPERIMENTAL STUDIES

Author	Title	Type	N	Subjects	Incentives	Elicitation measures
Aharoni and Fridlund, 2007	“Social reactions toward people vs. computers: how mere labels shape interactions”	Online	40	Students	Non-contingent	Stated, revealed preferences
Ahmad et al., 2009	“Computer-assisted screening for intimate partner violence and control: a randomized trial”	RCT	144 (treatment) + 149 (control)	Adult women	N/A	Stated, revealed preferences.
Alberdi, Povyakalo, Strigini, and Ayton, 2004	“Effects of incorrect computer-aided detection (CAD) output on human decision-making in mammography”	1: Lab in the field	20	Clinical experts	N/A	Stated, revealed preferences
		2: Lab in the field	19	Clinical experts	N/A	Stated, revealed preferences
Alberdi, Povyakalo, Strigini, Ayton, and Given-Wilson, 2008	“CAD in mammography: lesion-level versus case-level analysis of the effects of prompts on human decisions”	Lab in the field	50	Clinical experts	N/A	Revealed preferences
Arkes et al., 2007	“Patients derogate physicians who use a computer-assisted diagnostic aid”	1: Lab	347	Students	Non-contingent	Stated preferences

Table 1: EXPERIMENTAL STUDIES

Author	Title	Type	N	Subjects	Incentives	Elicitation measures
		2: Lab	128	Students	Non-contingent	Stated preferences
		3: Field	74	Hospital patients	Non-contingent	Stated preferences
		4: Lab	131	Medical students	Non-contingent	Stated preferences
ω Bahner et al., 2008	“Misuse of diagnostic aids in process control: the effects of automation misses on complacency and automation bias”	Lab	24	Students	Non-contingent	Revealed preferences
Bai et al., 2020	“The impacts of algorithmic work assignment on fairness perceptions and productivity: evidence from field experiments”	1: Field	50	Warehouse workers	Non-contingent	Stated, revealed preferences
		Replication: Field	20	Warehouse workers	Non-contingent	Stated, revealed preferences
Bartneck, Rosalia, et al., 2005	“Robot abuse – a limitation of the media equation”	Lab	20	Students & university employees	Non-contingent	Revealed preferences

Table 1: EXPERIMENTAL STUDIES

Author	Title	Type	N	Subjects	Incentives	Elicitation measures
Bartneck, Yogeeswaran, et al., 2018	“Robots and racism”	1: Online	192	Crowdfower participants	contingent	Stated, revealed preferences
		2: Online	172	Crowdfower participants	contingent	Stated, revealed preferences
Bickmore et al., 2005	“Establishing the computer – patient working alliance in automated health behavior change interventions”	Online	91	Adults	Non- contingent	Stated, revealed preferences
Bigman and K. Gray, 2018	“People are averse to machines making moral decisions”	1: Online	242	MTurk participants	Non- contingent	Stated preferences
		2: Online	241	MTurk participants	Non- contingent	Stated preferences
		3: Online	240	MTurk participants	Non- contingent	Stated preferences
		4: Online	242	MTurk participants	Non- contingent	Stated preferences
		5: Online	485	MTurk participants	Non- contingent	Stated preferences

Table 1: EXPERIMENTAL STUDIES

Author	Title	Type	N	Subjects	Incentives	Elicitation measures
		6: Online	239	MTurk participants	Non-contingent	Stated preferences
		7: Online	100	MTurk participants	Non-contingent	Revealed preferences
		8: Online	240	MTurk participants	Non-contingent	Stated preferences
		9 (within subjects): Online	201	MTurk participants	Non-contingent	Revealed preferences
		9 (between subjects): Online	482	MTurk participants	Non-contingent	Stated preferences
Briggs and Scheutz, 2014	“How robots can affect human behavior: Investigating the effects of robotic displays of protest and distress”	1: Lab	20	students	N/A	Stated, revealed preferences
		2: Lab	13	students	N/A	Stated, revealed preferences
		3: Lab	14	students	N/A	Stated, revealed preferences

Table 1: EXPERIMENTAL STUDIES

Author	Title	Type	N	Subjects	Incentives	Elicitation measures
Bundorf et al., 2019	“How do humans interact with algorithms? Experimental evidence from health insurance”	RCT	1,185	Patients	Non-contingent	Stated, revealed preferences
Castelo, 2019	“Blurring the Line Between Human and Machine: Marketing Artificial Intelligence”	Ch. 3: 1: Online	387	MTurk participants	N/A	Stated preferences
		Ch. 3: 2: Online	41,592 views, 604 clicks	Facebook users	None	Revealed preferences
		Ch. 3: 3: Online	201	MTurk participants	N/A	Stated preferences
		Ch. 3: 4: Online	201	Prolific academic participants	N/A	Stated preferences
		Ch. 3: 5: Online	13,621 views, 101 clicks	Facebook users	None	Revealed preferences

Table 1: EXPERIMENTAL STUDIES

Author	Title	Type	N	Subjects	Incentives	Elicitation measures
		Ch. 3: 6: Online	399	Prolific academic participants	Contingent	Stated, revealed preferences
		Ch 4.: 1: Online	800	US Prolific academic participants	N/A	Stated preferences
		Ch 4.: 2: Online	100	MTurk participants	N/A	Stated, revealed preferences
		Ch 4.: 3: Online	282	Prolific academic participants	N/A	Stated preferences
		Ch 4.: 4: Online	300	Prolific academic participants	N/A	Stated preferences
		Ch 4.: 5: Lab	83	Students	N/A	Stated preferences & physiological measures

Table 1: EXPERIMENTAL STUDIES

Author	Title	Type	N	Subjects	Incentives	Elicitation measures
Chaminade et al., 2012	“How do we think machines think? an fMRI study of alleged competition with an artificial intelligence”	Lab	19	Male students	Non-contingent	fMRI & stated preferences
Cohn et al., 2018	“Honesty in the digital age”	1: Online	486	Students	Contingent	Stated, revealed preferences
		2: Online	380	Students	Contingent	Stated, revealed preferences
∞ Corgnet et al., 2019	“Rac(g)e Against the Machine?: Social Incentives When Humans Meet Robots”	Lab	240	University-educated young adults	Contingent	Stated, revealed preferences
Coricelli and Nagel, 2009	“Neural correlates of depth of strategic reasoning in medial prefrontal cortex”	Lab	20	Adults	Contingent	fMRI & stated preferences
Cormier et al., 2013	“Would you do as a robot commands? An obedience study for human-robot interaction”	Lab	27	Adults	Non-contingent	Stated, revealed preferences
Cowgill, 2018	“Bias and productivity in humans and algorithms: Theory and evidence from resume screening”	Field	N/A	HR professionals	N/A	Revealed preferences
Cowgill et al., 2020	“The managerial effects of algorithmic fairness activism”	1: Online	ca. 500	US adults	N/A	Stated preferences

Table 1: EXPERIMENTAL STUDIES

Author	Title	Type	N	Subjects	Incentives	Elicitation measures
		2: Online	ca. 500	US adults	N/A	Stated preferences
De Laere et al., 1998	“The electronic mirror: human-computer interaction and change in self-appraisals”	Lab	158	Students	Non-contingent	Stated preferences
Melo, Carnevale, et al., 2011	“The effect of expression of anger and happiness in computer agents on negotiations with humans”	Lab	150	Students	Contingent	Stated, revealed preferences
Dietvorst and Bharti, 2019	“People reject algorithms in uncertain decision domains because they have diminishing sensitivity to forecasting error”	1: Online	601	MTurk participants	Contingent	Revealed preferences
		2: Online	403	MTurk participants	Contingent	Stated, revealed preferences
		3: Online	1,005	MTurk participants	Contingent	Stated, revealed preferences
		4: Online	405	MTurk participants	Contingent	Revealed preferences
		5a: Online	401	MTurk participants	Contingent	Revealed preferences

Table 1: EXPERIMENTAL STUDIES

Author	Title	Type	N	Subjects	Incentives	Elicitation measures
		5b: Online	399	MTurk participants	Contingent	Revealed preferences
Dietvorst, Simmons, et al., 2015	“Algorithm aversion: people erroneously avoid algorithms after seeing them err.”	1: Lab	361	N/A	Contingent	Stated, revealed preferences
		2: Lab	206	N/A	Contingent	Stated, revealed preferences
		3a: Online	410	MTurk participants	Contingent	Stated, revealed preferences
		3b: Online	1,036	MTurk participants	Contingent	Stated, revealed preferences
		4: Lab	354	N/A	Contingent	Stated, revealed preferences
Dietvorst, Simmons, et al., 2018	“Overcoming algorithm aversion: people will use imperfect algorithms if they can (even slightly) modify them”	1: Lab	288	N/A	Contingent	Stated, revealed preferences
		2: Online	816	MTurk participants	Contingent	Revealed preferences

Table 1: EXPERIMENTAL STUDIES

Author	Title	Type	N	Subjects	Incentives	Elicitation measures
		3: Online	818	MTurk participants	Contingent	Stated, revealed preferences
Dijkstra, 1999	“User agreement with incorrect expert system advice”	Lab	73	Students	Non-contingent	Stated preferences
Dijkstra et al., 1998	“Persuasiveness of expert systems”	Lab	85	Students	Non-contingent	Stated preferences
Domeinski et al., 2007	“Human redundancy in automation monitoring: effects of social loafing and social compensation”	Lab	36	Students	N/A	Stated, revealed preferences
Dzindolet et al., 2002	“The perceived utility of human and automated aids in a visual detection task”	1: Lab	68	Students	N/A	Stated, revealed preferences
		2: Lab	128	Students	Contingent	Stated, revealed preferences
		3: Lab	71	Students	Contingent	Stated, revealed preferences
Eyssel and Hegel, 2012	“(s)he’s got the look: Gender stereotyping of robots”	Lab	60	Students	N/A	Stated, revealed preferences

Table 1: EXPERIMENTAL STUDIES

Author	Title	Type	N	Subjects	Incentives	Elicitation measures
Farjam and Kirchkamp, 2018	“Bubbles in hybrid markets: how expectations about algorithmic trading affect human trading”	Lab	216	Students	Contingent	Revealed preferences
Fogg and Nass, 1997	“How users reciprocate to computers: an experiment that demonstrates behavior change”	Lab	76	N/A	N/A	Revealed preferences
Galletta et al., 2005	“Does spell-checking software need a warning label?”	Lab	65	Students	N/A	Revealed preferences
Galster et al., 2001	“Air traffic controller performance and workload under mature free flight: Conflict detection and resolution of aircraft self-separation”	Lab	10	Experts	None	Stated, revealed preferences
Goetz et al., 2003	“Matching robot appearance and behavior to tasks to improve human-robot cooperation”	1: Online	108	Students	N/A	Revealed preferences
		2: Lab	21	Students	N/A	Stated, revealed preferences
		3: Online	47	Students	N/A	Stated, revealed preferences

Table 1: EXPERIMENTAL STUDIES

Author	Title	Type	N	Subjects	Incentives	Elicitation measures
Gogoll and Uhl, 2018	“Rage against the machine: automation in the moral domain”	Lab	264	Students	Contingent	Stated, revealed preferences
Gombolay et al., 2015	“Decision-making authority, team efficiency and human worker satisfaction in mixed human–robot teams”	1: Lab	N/A	Students and young professionals	N/A	Stated, revealed preferences
		2: Lab	N/A	Students and young professionals	N/A	Stated, revealed preferences
Gratch et al., 2007	“Creating rapport with virtual agents”	Lab	131	Adults	Non-contingent	Stated, revealed preferences
H. M. Gray et al., 2007	“Dimensions of mind perception”	Online	2,399	Adults interested in social and moral psychology	N/A	Revealed preferences
K. Gray and Wegner, 2012	“Feeling robots and human zombies: mind perception and the uncanny valley”	1: Lab-in-the-field	120	Adults, recruited in subway stations and campus dining halls	N/A	Stated preferences

Table 1: EXPERIMENTAL STUDIES

Author	Title	Type	N	Subjects	Incentives	Elicitation measures
		2: Lab-in-the-field	45	Adults, recruited in subway stations and campus dining halls	N/A	Stated preferences
		2b : Online	28	MTurk participants	N/A	Stated preferences
		3: Lab-in-the-field	44	Adults, recruited in subway stations and campus dining halls	N/A	Stated preferences
Hertz and Wiese, 2019	“Good advice is beyond all price, but what if it comes from a machine?”	N/A	68	Students	Non-contingent	Revealed preferences
Hinds et al., 2004	“Whose job is it anyway? A study of human-robot interaction in a collaborative task”	Lab	292	Students	Non-contingent	Stated, revealed preferences

Table 1: EXPERIMENTAL STUDIES

Author	Title	Type	N	Subjects	Incentives	Elicitation measures
Humphreys et al., 2011	“Increasing discussions of intimate partner violence in prenatal care using Video Doctor plus Provider Cueing: a randomized, controlled trial”	RCT	50	Pregnant women with IPV risk	N/A	Stated preferences
Jago, 2019	“Algorithms and authenticity”	1: Online	175	Students	N/A	Stated preferences
		2: Online	401	MTurk participants	N/A	Stated preferences
		3: Online	200	MTurk participants	N/A	Stated preferences
		4: Online	804	MTurk participants	N/A	Stated preferences
Kantowitz et al., 1997	“Driver acceptance of unreliable traffic information in familiar and unfamiliar settings”	Lab	48	Young adults (able to drive)	Contingent	Stated, revealed preferences
Katagiri et al., 2001	“Cross-cultural studies of the computers are social actors paradigm: The case of reciprocity”	1: Lab	22+22	US and Japanese	N/A	Revealed preferences

Table 1: EXPERIMENTAL STUDIES

Author	Title	Type	N	Subjects	Incentives	Elicitation measures
		2: Lab	80	Japanese	N/A	Revealed preferences
Kirchkamp and Strobel, 2019	“Sharing responsibility with a machine”	Lab	399	Students	Contingent	Stated, revealed preferences
Kizilcec, 2016	“How much information? Effects of transparency on trust in an algorithmic interface”	Field	103	Individuals enrolled in an open online course	Non-contingent	Stated preferences
Krach et al., 2008	“Can machines think? interaction and perspective taking with robots investigated via fMRI”	Lab	20	Male adults	Non-contingent	fMRI & stated, revealed preferences
E.-J. Lee, 2010	“What triggers social responses to flattering computers? Experimental tests of anthropomorphism and mindlessness explanations”	1: Lab	204	Students	Contingent	Stated, revealed preferences
		2: Lab	149	Students	Contingent	Stated, revealed preferences

Table 1: EXPERIMENTAL STUDIES

Author	Title	Type	N	Subjects	Incentives	Elicitation measures
M. K. Lee, 2018	“Understanding perception of algorithmic decisions: fairness, trust, and emotion in response to algorithmic management”	Online	228	MTurk participants (US residents)	Non-contingent	Stated preferences
M. K. Lee and Baykal, 2017	“Algorithmic mediation in group decisions: Fairness perceptions of algorithmically mediated vs. discussion-based social division”	1: Lab	55	Participants recruited through a university-managed participant recruitment website	Non-contingent	Stated preferences
		2: Lab	103	Participants recruited through a university-managed participant recruitment website	Non-contingent	Stated preferences

Table 1: EXPERIMENTAL STUDIES

Author	Title	Type	N	Subjects	Incentives	Elicitation measures
M. K. Lee, Jain, et al., 2019	“Procedural justice in algorithmic fairness: leveraging transparency and outcome control for fair algorithmic mediation”	Lab	71	N/A	Non-contingent	Stated, revealed preferences
Leyer and Schneider, 2019	“Me, You or AI? How Do We Feel About Delegation”	Online	1,246	Students, university employees and others	N/A	Stated, revealed preferences
Lim and Reeves, 2010	“Computer agents versus avatars: responses to interactive game characters controlled by a computer or other player”	Lab	34	Students	N/A	Stated preferences & physiological measures
Logg et al., 2019	“Algorithm appreciation: People prefer algorithmic to human judgment”	1a: Online	202	MTurk participants	Contingent	Revealed preferences
		1b: Online	215	MTurk participants	Contingent	Revealed preferences
		1c: Online	286	MTurk participants	N/A	Revealed preferences
		1d: Online	119	Experts	Contingent	Revealed preferences

Table 1: EXPERIMENTAL STUDIES

Author	Title	Type	N	Subjects	Incentives	Elicitation measures
		2: N/A	154	N/A	N/A	Revealed preferences
		N/A	403	N/A	Contingent	Revealed preferences
		4: Online	70 experts, 301 MTurk participants	Experts and MTurk participants	Contingent	Revealed preferences
Longoni et al., 2019	“Resistance to medical artificial intelligence”	1: N/A	228	Students	Non-contingent	Revealed preferences
		2: Online	103	MTurk participants	Non-contingent	Stated preferences
		3 (A-C): Online	744	MTurk participants	Non-contingent	Stated preferences
		4: Online	100	MTurk participants	Non-contingent	Revealed preferences
		5: Online	286	MTurk participants	Non-contingent	Stated preferences

Table 1: EXPERIMENTAL STUDIES

Author	Title	Type	N	Subjects	Incentives	Elicitation measures
		6: Online	243	MTurk participants	Non-contingent	Stated preferences
		7: Online	294	MTurk participants	Non-contingent	Stated preferences
		8: Online	401	MTurk participants	Non-contingent	Stated preferences
		9: Online	197	MTurk participants	Non-contingent	Stated preferences
Lucas et al., 2014	“It’s only a computer: virtual humans increase willingness to disclose”	Lab	154	Adults	N/A	Stated, revealed preferences
Luo et al., 2019	“Frontiers: machines vs. humans: The impact of artificial intelligence chatbot disclosure on customer purchases”	Field	6200	Customers of a financial service company	None	Revealed preferences
Mandryk et al., 2006	“Using psychophysiological techniques to measure user experience with entertainment technologies”	1: Lab	7	Male students	N/A	Stated preferences & physiological measures

Table 1: EXPERIMENTAL STUDIES

Author	Title	Type	N	Subjects	Incentives	Elicitation measures
		2: Lab	10	Students	N/A	Stated preferences & physiological measures
Manzey et al., 2012	“Human performance consequences of automated decision aids: the impact of degree of automation and system experience”	1: Lab	56	Students	Non-contingent	Stated, revealed preferences
		2: Lab	88	Students	Non-contingent	Stated, revealed preferences
McCabe et al., 2001	“A functional imaging study of cooperation in two-person reciprocal exchange”	Lab	12	N/A	Contingent	fMRI
Melo, Marsella, et al., 2016	“People do not feel guilty about exploiting machines”	1: Online	81	MTurk participants	Contingent	Stated, revealed preferences
		2: N/A	165	Students	Contingent	Revealed preferences
Mende et al., 2019	“Service robots rising: how humanoid robots influence service experiences and elicit compensatory consumer responses”	1a: Lab	80	Students	Non-contingent	Revealed preferences

Table 1: EXPERIMENTAL STUDIES

Author	Title	Type	N	Subjects	Incentives	Elicitation measures
		1b: Lab	253	Students	Non-contingent	Revealed preferences
		1c: Lab	215	Students	Non-contingent	Revealed preferences
		2: Online	100	MTurk participants	N/A	Stated, revealed preferences
		3a: Online	180	MTurk participants	N/A	Revealed preferences
		3b: Lab	203	Students	Non-contingent	Revealed preferences
		4: Lab	250	Students	Non-contingent	Revealed preferences
Metzger and Parasuraman, 2005	“Automation in future air traffic management: effects of decision aid reliability on controller performance and mental workload”	1: Lab	12	Experts	Non-contingent	Stated, revealed preferences
		2: Lab	12	Experts	Non-contingent	Stated, revealed preferences

Table 1: EXPERIMENTAL STUDIES

Author	Title	Type	N	Subjects	Incentives	Elicitation measures
Metzger, Duley, et al., 2000	“Effects of variable-priority training on automation-related complacency: performance and eye movements”	Lab	43	Students	Non-contingent	Revealed preferences & eye movement data
Mirnig et al., 2017	“To err is robot: How humans assess and act toward an erroneous social robot”	Lab	45	Adults	N/A	Stated, revealed preferences
Moon, 2000	“Intimate exchanges: using computers to elicit self-disclosure from consumers”	1: Lab	60	Students	Non-contingent	Stated, revealed preferences
		2: Lab	24	Students	Non-contingent	Stated, revealed preferences
Moon, 2003	“Don’t blame the computer: when self-disclosure moderates the self-serving bias”	1: Lab	48	Adults	Non-contingent	Stated, revealed preferences
		2: Lab	62	Adults	Non-contingent	Stated, revealed preferences
Moon and Nass, 1998	“Are computers scapegoats? Attributions of responsibility in human – computer interaction”	Lab	80	Students	Non-contingent	Stated preferences

Table 1: EXPERIMENTAL STUDIES

Author	Title	Type	N	Subjects	Incentives	Elicitation measures
Mosier, Palmer, et al., 1992	“Electronic checklists: implications for decision making”	Lab	24	Experts	N/A	Stated, revealed preferences
Mosier, Skitka, Heers, et al., 1998	“Automation bias: decision making and performance in high-tech cockpits”	Lab	25	Experts	N/A	Stated, revealed preferences
Mosier, Skitka, Dunbar, et al., 2001	“Aircrews and automation bias: the advantages of teamwork?”	Lab	48	Commercial pilots	N/A	Stated, revealed preferences
Muir and Moray, 1996	“Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation”	1: Lab	6	Students (Male)	Contingent	Stated, revealed preferences
		2: Lab	6	Students (Male)	Contingent	Stated, revealed preferences
Kurtzberg and Naquin, 2004	“Human reactions to technological failure: How accidents rooted in technology vs. human error influence judgments of organizational accountability”	1: Lab	86	Students	N/A	Stated preferences
		2: Lab	89	Students	N/A	Stated preferences

Table 1: EXPERIMENTAL STUDIES

Author	Title	Type	N	Subjects	Incentives	Elicitation measures
Nass, Fogg, et al., 1996	“Can computers be teammates?”	Lab	56	Students	N/A	Stated, revealed preferences
Nass, Green, et al., 1997	“Are machines gender neutral? Gender-stereotypic responses to computers with voices”	Lab	40	Students	N/A	Stated preferences
Önkal et al., 2009	“The relative influence of advice from human experts and statistical methods on forecast adjustments”	1: Lab	76	Students	Non-contingent	Stated, revealed preferences
		2: Lab	54	Students	Non-contingent	Stated, revealed preferences
Palmeira and Spassova, 2015	“Consumer reactions to professionals who use decision aids”	1: Online	70	US Adults	Non-contingent	Stated preferences
		2: Online	192	US Adults	Non-contingent	Stated preferences
		3: Online	83	US Adults	Non-contingent	Stated preferences
Parasuraman, Molloy, et al., 1993	“Performance consequences of automation-induced ‘complacency’”	1: Lab	24	Adults	Non-contingent	Revealed preferences

Table 1: EXPERIMENTAL STUDIES

Author	Title	Type	N	Subjects	Incentives	Elicitation measures
		2: Lab	16	Adults	Non-contingent	Revealed preferences
Paravisini and Schoar, 2013	“The incentive effect of scores: randomized evidence from credit committees”	RCT	1421	Credit committee members of a bank	N/A	Revealed preferences
M. V. Pezzo and S. P. Pezzo, 2006	“Physician evaluation after medical errors: does having a computer decision aid help or hurt in hindsight?”	1: N/A	59	Students	Non-contingent	Stated preferences
		2: N/A	320	Medical and non-medical students	Non-contingent	Stated preferences
Prahl and Van Swol, 2017	“Understanding algorithm aversion: when is advice from automation discounted?”	Online	157	Students	Non-contingent	Stated, revealed preferences
Promberger and Baron, 2006	“Do patients trust computers?”	1: Online	86	Adults	Non-contingent	Stated preferences
		2: Online	80	Adults	Non-contingent	Stated preferences

Table 1: EXPERIMENTAL STUDIES

Author	Title	Type	N	Subjects	Incentives	Elicitation measures
Ribeiro et al., 2016	“” Why should i trust you?” Explaining the predictions of any classifier”	1: Online	400	MTurk participants (with basic knowledge about religion)	N/A	Revealed preferences
		2: Online	10	MTurk participants	N/A	Stated, revealed preferences
Rosenthal-von der Pütten et al., 2013	“An experimental study on emotional reactions towards a robot”	Lab	18	Students	Non-contingent	Stated, revealed preferences
Rovira et al., 2007	“Effects of imperfect automation on decision making in a simulated command and control task”	Lab	18	Students	Non-contingent	Stated, revealed preferences
Salem et al., 2015	“Would you trust a (faulty) robot? Effects of error, task type and personality on human-robot cooperation and trust”	Lab	40	Adults	N/A	Stated, revealed preferences
Salomons et al., 2018	“Humans conform to robots: Disambiguating trust, truth, and conformity”	Lab	30	Students	Non-contingent	Stated, revealed preferences

Table 1: EXPERIMENTAL STUDIES

Author	Title	Type	N	Subjects	Incentives	Elicitation measures
Sanfey et al., 2003	“The neural basis of economic decision – making in the ultimatum game”	Lab	19	N/A	Contingent	Revealed preferences & fMRI
Sarter and Schroeder, 2001	“Supporting decision making and action selection under time pressure and uncertainty: the case of in-flight icing”	Lab	27	Commercial pilots	Non-contingent	Revealed preferences
Saygin et al., 2012	“The thing that should not be: predictive coding and the uncanny valley in perceiving human and humanoid robot actions”	Lab	20	Adults	N/A	fMRI
Schniter et al., 2020	“Trust in humans and robots: economically similar but emotionally different”	Lab	387	Students	Contingent	Stated, revealed preferences
Shaffer et al., 2013	“Why do patients derogate physicians who use a computer-based diagnostic support system?”	1: N/A	434	Students	Non-contingent	Stated preferences
		2: Online	109	Students	Non-contingent	Stated preferences
		3: N/A	189	Students	Non-contingent	Stated preferences

Table 1: EXPERIMENTAL STUDIES

Author	Title	Type	N	Subjects	Incentives	Elicitation measures
Shank, 2012	“Perceived justice and reactions to coercive computers”	Lab	114	Students	Contingent	Stated, revealed preferences
Short, Hart, et al., 2010	“No fair!! an interaction with a cheating robot”	Lab	60	Students	N/A	Stated, revealed preferences
Short and Mataric, 2017	“Robot moderation of a collaborative game: Towards socially assistive robotics in group interactions”	Lab	30	Freshman IT students	N/A	Stated, revealed preferences
29 Skitka, Mosier, and Burdick, 1999	“Does automation bias decision-making?”	Lab	80	Students	Non-contingent	Stated, revealed preferences
Skitka, Mosier, and Burdick, 2000	“Accountability and automation bias”	Lab	181	Students	Non-contingent	Stated, revealed preferences
Skitka, Mosier, Burdick, and Rosenblatt, 2000	“Automation bias and errors: are crews better than individuals?”	Lab	144	Students	Non-contingent	Revealed preferences
Slater et al., 2006	“A virtual reprise of the Stanley Milgram obedience experiments”	Lab	34	Students & University Employees	N/A	Stated, revealed preferences & physiological measures

Table 1: EXPERIMENTAL STUDIES

Author	Title	Type	N	Subjects	Incentives	Elicitation measures
Sproull et al., 1996	“When the interface is a face”	Lab	130	Students	Non-contingent	Stated, revealed preferences
Strohkorb et al., 2016	“Improving human-human collaboration between children with a social robot”	Lab	86	Children (6-9 y.o.)	N/A	Stated, revealed preferences
Tay, Park, et al., 2013	“When stereotypes meet robots: the effect of gender stereotypes on people’s acceptance of a security robot”	Lab	40	Students	Non-contingent	Stated preferences
Tay, Jung, et al., 2014	“When stereotypes meet robots: the double-edge sword of robot gender and personality in human-robot interaction”	Lab	164	Students	N/A	Stated, revealed preferences
Tan et al., 2018	“Investigating human+ machine complementarity for recidivism predictions”	Online	20	MTurk participants	N/A	Stated preferences
Tazelaar and Snijders, 2004	“The myth of purchasing professionals’ expertise. More evidence on whether computers can make better procurement decisions”	Original: N/A	91	Students & purchasing professionals	N/A	Stated preferences
		I&L test - 1: online	72	Mostly purchasing managers	N/A	Stated preferences

Table 1: EXPERIMENTAL STUDIES

Author	Title	Type	N	Subjects	Incentives	Elicitation measures
		I&L test - 2: online	72	Mostly purchasing managers	N/A	Stated preferences
		The Masterclass: lab	13	Purchasing professionals	N/A	Stated preferences
		Replication and extension: N/A	118	Students & purchasing professionals & people with no knowledge about purchasing	N/A	Stated preferences
Teubner et al., 2015	“The impact of computerized agents on immediate emotions, overall arousal and bidding behavior in electronic auctions”	Lab	120	Students	Contingent	Revealed preferences & physiological measures
Traeger et al., 2020	“Vulnerable robots positively shape human conversational dynamics in a human–robot team”	Lab	153	N/A	non-contingent	revealed and stated

Table 1: EXPERIMENTAL STUDIES

Author	Title	Type	N	Subjects	Incentives	Elicitation measures
Tsai et al., 2003	“Computer decision support as a source of interpretation error: the case of electrocardiograms”	RCT	30	Physicians	N/A	Revealed preferences
Wout et al., 2006	“Affective state and decision-making in the ultimatum game”	Lab	30	Students	Contingent	Revealed preferences & physiological measures
Von der Puetten et al., 2010	““It doesn’t matter what you are!” explaining social effects of agents and avatars.”	Lab	83	Adults	Non-contingent	Revealed preferences
Waytz and Norton, 2014	“Botsourcing and outsourcing: robot, British, Chinese, and German workers are for thinking — not feeling — jobs.”	1: Online	103	MTurk participants	Non-contingent	Stated preferences
		2: Online	266	MTurk participants	Non-contingent	Stated preferences
		3: Online	54	MTurk participants	Non-contingent	Stated preferences
		4: Online	153	MTurk participants	Non-contingent	Stated preferences

Table 1: EXPERIMENTAL STUDIES

Author	Title	Type	N	Subjects	Incentives	Elicitation measures
		5: Online	167	MTurk participants	Non-contingent	Stated preferences
		6a: Online	166	MTurk participants	Non-contingent	Stated preferences
		6b: Online	167	MTurk participants	Non-contingent	Stated preferences
		6c: Online	164	MTurk participants	Non-contingent	Stated preferences
Waytz, Heafner, et al., 2014	“The mind in the machine: anthropomorphism increases trust in an autonomous vehicle”	Lab	100	Adults (able to drive)	N/A	Stated preferences & physiological measures
Wickens et al., 2015	“Complacency and automation bias in the use of imperfect automation”	Lab	119	Students	Non-contingent	Revealed preferences
Yeomans et al., 2019	“Making sense of recommendations”	1a: N/A	122	Museum visitors	N/A	Revealed preferences
		1b: Online	544	MTurk participants	Contingent	Revealed preferences

Table 1: EXPERIMENTAL STUDIES

Author	Title	Type	N	Subjects	Incentives	Elicitation measures
		2: N/A	210	Museum visitors	N/A	Revealed preferences
		3: Online	886	MTurk participants	N/A	Stated, revealed preferences
		4: Online	899	MTurk participants	N/A	Stated, revealed preferences
		5: Online	972	MTurk participants	N/A	Stated, revealed preferences
Zhang et al., 2010	Lab in the field	“Service robot feature design effects on user perceptions and emotional responses”	24	Seniors (64-91 y.o.)	N/A	Stated, revealed measures

Table 2: OBSERVATIONAL STUDIES

Author	Title	Data	Topic
Broek, 2019	“Hiring algorithms: an ethnography of fairness in practice”	Large European FMCG company	Use of an AI tool in hiring decisions
Gates et al., 2002	“Automated underwriting in mortgage lending: good news for the underserved?”	Federal Home Loan Mortgage Corporation	Use of automated underwriting systems in mortgage lending decisions
Hoffman et al., 2018	“Discretion in hiring”	Personnel data of 15 firms in service sector with low-skilled workers	Identification of human biases in hiring by comparison of actual hiring decisions with automated recommendations
Kleinberg et al., 2018	“Human decisions and machine predictions”	Arrests in New York City	Use of machine learning in judicial recidivity risk decisions
Lebovitz et al., 2019	“Doubting the diagnosis: how artificial intelligence increases ambiguity during professional decision making”	Department of radiology in a US hospital	Use of AI tools in diagnostic decision-making in radiology
Saez de Tejada Cuenca, 2019	“Essays on Social and Behavioral Aspects of Apparel Supply Chains”	Large clothing retailer	Incorporation of decision aid support regarding price-setting in sales decisions by managers
Stevenson, 2018	“Assessing risk assessment in action”	Criminal cases in Kentucky	Use of an algorithmic risk assessment tool in the criminal justice system
Stubbs et al., 2007	“Autonomy and common ground in human-robot interaction: a field study”	Observations of scientific field work (Life in the Atacama project)	Use of robot to support scientific research

Table 3: LITERATURE REVIEWS

Author	Title	Topic
Alberdi, Strigini, et al., 2009	“Why are people’s decisions sometimes worse with computer support?”	Automation bias and mechanisms involving human errors when using computer support
Burton et al., 2020	“A systematic review of algorithm aversion in augmented decision making”	Algorithm aversion with the focus on conditions that lead to the acceptance or rejection of algorithmically generated insights
Glikson and Woolley, 2020	“Human trust in Artificial Intelligence: Review of empirical research”	The determinants of human trust in AI

Table 3: LITERATURE REVIEWS

Author	Title Topic	
Goddard et al., 2012	“Automation bias: a systematic review of frequency, effect mediators, and mitigators”	Automation bias with focus on frequency, effect mediators, and mitigators
Jussupow et al., 2020	“Why are we averse towards Algorithms? A comprehensive literature Review on Algorithm aversion.”	Algorithm aversion and algorithm appreciation
J. D. Lee and See, 2004	“Trust in automation: designing for appropriate reliance”	Link between trust and reliance on automation
March, 2019	“The behavioral economics of artificial intelligence: Lessons from experiments with computer players”	Review of experimental studies that use computer players
Mende et al., 2019	“Service robots rising: how humanoid robots influence service experiences and elicit compensatory consumer responses”	Use of robots in service settings (part of the paper)
Mosier and Skitka, 1996	“Human decision makers and automated decision aids: made for each other?”	Human interactions with automated decision aids
Nass and Moon, 2000	“Machines and mindlessness: social responses to computers”	Summary of a series of experimental studies that demonstrate that individuals mindlessly apply social rules and expectations to computers
Parasuraman and Riley, 1997	“Humans and automation: use, misuse, disuse, abuse”	Theoretical, empirical, and analytical studies pertaining to human use, misuse, disuse, and abuse of automation technology
Parasuraman and Manzey, 2010	“Complacency and bias in human use of automation: an attentional integration”	Complacency and bias in interactions with automated decision support systems
Złotowski et al., 2015	“Anthropomorphism: opportunities and challenges in human–robot interaction”	Summary of potential benefits and challenges of building anthropomorphic robots, from both a philosophical perspective and from the viewpoint of empirical research in the fields of human–robot interaction and social psychology