Computers in Human Behavior 124 (2021) 106879

Contents lists available at ScienceDirect

Computers in Human Behavior

journal homepage: www.elsevier.com/locate/comphumbeh

To err is human, not algorithmic – Robust reactions to erring algorithms

Laetitia A. Renier *, Marianne Schmid Mast, Anely Bekbergenova

Faculty of Business and Economics (HEC), University of Lausanne, Quartier Unil-Chamberonne, CH-1015, Lausanne, Switzerland

ARTICLE INFO

Keywords:
Algorithm aversion
Artificial intelligence
Error
Reactions
Perception
Third-party

ABSTRACT

When seeing algorithms err, we trust them less and decrease using them compared to after seeing humans err; this is called *algorithm aversion*. This paper builds on the algorithm aversion literature and the third-party reactions to mistreatment model to investigate a wider array of reactions to erring algorithms. Using an experimental design deployed with a vignette-based online study, we investigate gut reactions, justice cognitions, and behavioral intentions toward erring algorithms (compared to erring humans). Our results show that when the error was committed by an algorithm (vs. a human), gut reactions were harsher (i.e., less acceptance and more negative feelings), justice cognitions weaker (i.e., less blame, less forgiveness, and less accountability), and behavioral intentions stronger. These results remain independent of factors such as the maturity of the algorithms (better than or same as human performance), the severity of the error (high or low), and the domain of use (recruitment or finance). We discuss how these results complement the current literature thanks to a robust and more nuanced pattern of reactions to erring algorithms.

1. Introduction

Algorithm-based decision makers are an integral part of human life. They are not only used to accomplish tasks more quickly and efficiently than humans, they are also used to reduce human error (Lee, Nagy, Weaver, & Newman-Toker, 2013; Patel et al., 2010), to prevent bias in human decision making (Miller, 2018), or to assist humans in decision making. However, algorithms (e.g., mathematical calculations or artificial intelligence), like humans, are not error-free and can be biased (Dastin, 2018, pp. 5–9; Lambrecht & Tucker, 2019). Errors become particularly critical when decisions concern humans directly, such as deciding who obtains a mortgage loan (Markus, Dutta, Steinfield, & Wigand, 2008; Straka, 2000) or who is hired for a job (Upadhyay & Khandelwal, 2018). While erring is considered human, it is less acceptable when algorithms err because we expect imperfection only from humans; automation is supposed to be perfect (Madhavan & Wiegmann, 2007). Indeed, research on algorithm aversion has shown that we are less likely to trust or rely on algorithms compared to trusting or relying on humans after they have made errors (Dietvorst, Simmons, & Massey, 2014, 2018; Dzindolet, Peterson, Pomranky, Pierce, & Beck, 2003; Prahl & Van Swol, 2017).

The goal of the present research is to build on the literature of algorithm aversion and to extend it by investigating (a) a wider array of reactions (i.e., gut reactions in the form of acceptance and negative emotions; justice cognitions comprising blame, forgiveness, and perceived accountability, as well as behavioral intentions) after having learned about an erring algorithmic compared to an erring human decision maker and (b) key boundary conditions of algorithm aversion (i.e., algorithm maturity, severity of the error, and domain of use). We apply a strict experimental protocol in order to test the single and joint effects of these factors on the differences in reactions to algorithmic compared to human error.

This study contributes to the field of algorithm aversion in different ways. We look at a broader array of reactions (i.e., emotional, cognitive, and behavioral) to erring algorithms in one and the same study by using the third-party reactions to mistreatment model (O'Reilly & Aquino, 2011). The term algorithm averion connotes the negativity of the reactions toward erring algorithms. We set out to test to what extent the reactions show that algorithms are perceived and treated as "non-human" rather than simply in a more negative way than erring humans. Obtaining a more fine-grained picture of how exactly erring algorithms are perceived and reacted upon compared to humans is important for being able to communicate efficiently about the potential for algorithm use and for legislators who face challenges in assigning legal responsibility in case of erring algorithms.

Algorithms can be a great help for decision making, but their

* Corresponding author. Internef #558, Department of Organizational Behavior, Faculty of Business and Economics (HEC), University of Lausanne, CH-1015, Lausanne. Switzerland.

E-mail addresses: laetitia.renier@unil.ch (L.A. Renier), marianne.schmidmast@unil.ch (M. Schmid Mast), anely.bekbergenova@unil.ch (A. Bekbergenova).

https://doi.org/10.1016/j.chb.2021.106879
Received 19 November 2020; Received in revised form 5 February 2021; Accepted 25 May 2021
Available online 30 May 2021

adoption oftentimes encounters resistance. It is therefore important to know the boundary conditions of the reactions toward erring algorithms. Some of the potential moderators have been studied already (see, for example, Alexander, Blinder, & Zak, 2018, maturity and the extent to which it is used by others, Brooks, Begum, & Yanco, 2016, for severity of the error; Castelo, Bos, & Lehmann, 2019, for domain of use and maturity). However, to our knowledge, no study has addressed these factors in one and the same study. Doing this enables us to compare the factors and test their relative, separate, and joint influence on reactions to erring algorithms.

1.1. Reactions to erring algorithms

How do we react when algorithms make a mistake? Research has studied the extent to which individuals rely on algorithm-based advice as compared to relying on human-based advice in cases in which the advice is erroneous (see Burton, Stein, & Jensen, 2020; Jussupow, Benbasat, & Heinzl, 2020, for reviews). Individuals typically rely less on erring algorithms than on erring humans (i.e., algorithm aversion, Dietvorst et al., 2014; Dzindolet et al., 2003; Prahl & Van Swol, 2017). Prahl and Van Swol (2017) tested algorithm aversion in an experimental setting in which participants were asked to complete 14 trials of forecasting related to medical operation management while receiving either computer-generated or human-generated advice (e.g., estimating the mean time of surgical operation based on charts and data while incentives are aligned with the accuracy of the forecasts). Participants received good advice (close to the actual forecast) until the sixth trial, when all participants received bad advice (far from the actual forecast). Results showed that after receiving bad advice, participants who received computer-generated advice showed a greater decrease in advice utilization than participants who received the same bad advice, but from a human. Thus, people relied less on algorithms in comparison to humans, after seeing them err.

Specific expectations about human and algorithm performance best explain algorithm aversion (Burton et al., 2020; Dietvorst et al., 2014; Dzindolet et al., 2003; Prahl & Van Swol, 2017). People expect humans to be imperfect and algorithms to be perfect (Madhavan & Wiegmann, 2007). Consequently, if a human makes an error, this is perceived as being normal and to be expected, and when it happens, judgment is rather lenient. If an algorithm makes an error, this is unexpected and contrary to its very nature of being perfect, which is why the reactions are more negative (see Dietvorst et al., 2014; Prahl & Van Swol, 2017 for similar explanations of the algorithm aversion phenomenon). In sum, research suggests that people have higher performance expectations toward algorithms than humans. Thus, witnessing an algorithm erring creates a higher violation of expectation than witnessing an erring human. This explains why individuals are more tolerant toward erring humans than erring algorithms and why there is a greater decrease in the utilization of algorithm-based advice than human-based advice, after seeing them err.

Research on algorithm aversion has so far focused on a few behavioral reactions to algorithmic error. In the present research, we enlarge this perspective and investigate different levels of reactions toward errors: gut reactions, justice cognitions, and behavioral intentions. We use the third-party reactions to mistreatment model, developed by O'Reilly and Aquino (2011), to identify types of reaction to erroneous algorithmic decisions. The third-party reactions to mistreatment model (O'Reilly & Aquino, 2011) advances that when individuals witness mistreatment, third-party reactions will take three forms. People will first react on an intuitive, gut level with an *intuition of moral violation* and *moral anger*. They will then engage in a more conscious cognitive assessment of the situation and engage in *justice cognition*. Finally, this will guide their behavior as *behavioral intentions* will form.

We suggest that individuals' reactions to an erring algorithm might also take those forms. First, when witnessing the error (made by an algorithm or by a human) and its consequences, individuals will automatically and intuitively judge the situation as good or bad, and they will be more or less outraged by it (i.e., intuition of moral violation and moral anger, O'Reilly & Aquino, 2011). In the case of our study, these *gut reactions* comprise *acceptance* and *emotional reactions* (i.e., anger, disgust, or hostility toward the decision maker who wronged someone). We expect that individuals will intuitively be less tolerant toward the erring algorithm than toward the erring human as demonstrated by the literature on algorithm aversion (see Dietvorst et al., 2014; Dzindolet et al., 2003; Prahl & Van Swol, 2017) and as suggested by the perfect automation schema (Madhavan & Wiegmann, 2007). Formally stated, we hypothesize that when learning about an error committed by an algorithm, participants' *gut reactions* are harsher (i.e., less acceptance and more negative emotions) than when learning about the same error committed by a human (Hypothesis 1).

According to the third-party reactions to mistreatment model (O'Reilly & Aquino, 2011), after the gut reaction to witnessing an error and its consequences, individuals will start to consciously assess the situation, engage in more deliberate cognition and therefore develop more elaborate insights into what happened, including moral judgment or *justice cognition* (O'Reilly & Aquino, 2011). In the present research, justice cognitions comprise *blame*, *forgiveness*, and *perceived accountability*. Justice cognitions might apply less to algorithms and research indeed shows that blame and forgiveness apply more to humans than to machines (Malle, Guglielmo, & Monroe, 2014; Pizarro, 2014). More particularly, Pizarro (2014) suggests that machines are not agentic entities, that they are less in control, less responsible, and lack intentionality. Therefore, when a machine makes an erroneous decision, people might not think of it in terms of blame or forgiveness in the same way that they would if a human made a mistake. Hence, even if humans tend to attribute intentionality to non-human entities (Pizarro, 2014), erring algorithms would be blamed less than erring humans. In a similar vein, organizations are held less accountable when failure is said to be technological rather than human (Naquin & Kurtzberg, 2004). Moreover, individuals preferred human-based as compared to algorithm-based medical recommendations, one reason being that patients can more easily shift responsibility to a human than to an algorithm (Promberger & Baron, 2006). In sum, algorithms might not be blamed, forgiven, and held accountable as much as humans. Justice cognitions might therefore be attributed less to erring algorithms than to erring humans. Formally stated, we hypothesize that when learning about an error committed by an algorithm, participants' *justice cognitions* are weaker (i.e., less blame, less forgiveness, less accountability) than when learning about the same error committed by a human (Hypothesis 2).

The third-party reactions to mistreatment model (O'Reilly & Aquino, 2011) also includes predictions about how individuals intend to behave when witnessing mistreatment: they can either do nothing, directly or indirectly punish the perpetrator, or help the victim (O'Reilly & Aquino, 2011). In the context of our study, there are four types of behavioral intentions. Individuals might want to improve the erring algorithm (train the human), stop using it (fire the human), or do nothing. Moreover, we also look at whether people intend to keep using the company that uses the algorithm (employs the human). We expect that there will be more motivation to act when confronted with an erring algorithm than an erring human. This is because believing that "to err is human" might not be a driving factor in taking action against the erring human and believing that algorithms should not err might motivate people to act toward the erring algorithm. Formally stated, we hypothesize that when learning about an error made by an algorithm, participants' *behavioral intentions* are to act (i.e., improve/train, stop using/fire, nothing can be done, or keep using), more so than when learning about the same error made by a human (Hypothesis 3).

1.2. Algorithm maturity: how well do algorithms perform?

Algorithms outperform humans at playing certain games (e.g., chess; Campbell, Hoane, & Hsu, 2002; go, Silver et al., 2016; video games such

as Starcraft, Vinyals et al., 2019), making certain medical decisions (e.g., skin lesions, Tschandl et al., 2019; surgical audits, Brzezicki et al., 2020), or reading comprehension (Rajpurkar et al., 2016; SQuAD, n.d.). Conversely, humans outperform algorithms at recognizing faces in videos (Phillips & O'Toole, 2014) or specific language processing tasks (e.g., sentence completion, Radford, Wu, Amodei, et al., 2019; 2019b).

Similar to humans, algorithms can also make errors or be biased. To illustrate this, in 1988, St George's Hospital Medical School was found guilty of discrimination by the UK Commission for Racial Equality because their computer program, designed to help screen applicants for job interviews, reproduced an existing gender bias by selecting less female than male applicants (Lowry & Macpherson, 1988). Thirty years later, Amazon had to stop using a recruiting algorithm because it taught itself (based on training data) to discriminate against women by excluding them during the recruitment process (Dastin, 2018, pp. 5–9). Finally, algorithms used to reduce the cost of job advertisements have been found to show opportunities less often to women than to men (Lambrecht & Tucker, 2019).

Clearly, whether algorithms perform better than humans depends on the task at hand and how well the algorithm has been trained. However, users do not necessarily know the performance level of the algorithm they are using; they may have exaggerated trust and positive performance expectations toward algorithms or, conversely, rely too little on algorithms (Hoff & Bashir, 2015; Parasuraman & Riley, 1997). We propose that receiving information about how mature algorithms are (e.g., whether they perform better than humans or similar to humans) affects humans' reactions toward erring algorithms. Being told that algorithms and humans have the same level of performance for a given task might dampen the expectation that algorithms are perfect and make users more tolerant toward an erring algorithm. Indeed, presenting an algorithm as being more human-like improved reliance on algorithms (Castelo et al., 2019). In contrast, affirming that algorithms perform better than humans might lead to less lenient attitudes toward an erring algorithm (i.e., harsher gut reactions and more intentions to act), and because such an algorithm would be perceived as less human-like, justice cognitions might be weaker. Formally stated, we hypothesize that when affirming that algorithms perform better than humans (compared to affirming that algorithms perform at the same level as humans), learning about an error committed by an algorithm and its consequences, participants' (a) *gut reactions* are harsher (i.e., less acceptance and more negative emotions), (b) *justice cognitions* are weaker (i.e., less blame, less forgiveness, and less accountability), and (c) *behavioral intentions* indicate a willingness to act (i.e., more willing to improve and to stop using, less convinced that nothing can be done and to keep using; Hypothesis 4).

1.3. Error severity and domain of use

We also look at whether the severity of the error, defined as how negatively the user is affected by the error, influences our predictions. Studies on mistreatment (e.g., child abuse, car accidents, rape, and malpractice) show that the higher the perceived severity of the outcome of the mistreatment, the more responsibility is attributed to the perpetrator (Robbennolt, 2000). However, none of the mistreatment cases concerned algorithms. Research in robotics shows that reactions toward robots (e.g., satisfaction, trust) are affected significantly by the severity of the robot's degree of failure (Brooks et al., 2016) in that more severe mistakes entail more negative reactions. We do not promote any specific hypotheses for the present study, but test whether error severity is a factor that moderates Hypotheses 1 to 4.

In terms of the domain of use, we focus on applicant screenings for job positions (i.e., recruitment) and mortgage loans (i.e., finance). In recruitment, automated screening ranks and preselects candidates for further recruitment (Upadhyay & Khandelwal, 2018). In finance, automated underwriting is used to avoid mortgage default risk (Markus et al., 2008). Algorithmic decisions in these two assessment domains are not always error- or bias-free (see previous section) and one might be seen as more mature than the other. To our knowledge, no research has compared perceived algorithm performance (and maturity) in recruitment and finance.

Automated loan underwriting has been used since the 1990s (Markus et al., 2008; Straka, 2000) while automated applicant screening is more recent (Derous & De Fruyt, 2016; van Esch, Black, & Ferolie, 2019). Additionally, automated loan underwriting relies more on objective data (e.g., income) than automated screening (e.g., personality traits). Errors or biased decisions in recruitment seem to be mediatized more than automation in finance (see examples in the previous section). Given their more recent use and thus novelty in recruitment, the lack of transparency in algorithms used for recruitment, and the over-mediatization of algorithmic errors in this domain might challenge the adoption of algorithms (Shariff, Bonnefon, & Rahwan, 2017). In sum, the use of algorithms in recruitment might be perceived as less mature than their use in finance. In this vein, we expect that algorithms used to perform recruitment tasks will be perceived as less mature than algorithms used to perform financial tasks (consistent with Castelo and colleagues' research, 2019).

The maturity of the algorithm appears to be linked to the field of use and to the specific task at hand. Therefore, algorithm maturity and domain of use are confounded. This is why in the present research we manipulate the domain of use (recruitment and finance) as well as the information about algorithm performance (maturity) independently from each other in order to test their separate or joint effects on reactions to a decision maker error. We explore whether the domain of use is a factor that potentially moderates Hypotheses 1 to 4.

2. Method

2.1. Participants

We recruited 880 participants (439 women, $M = 36.34$, $SD = 12.26$) via the Prolific platform for a 15-min online experiment, remunerated with £1.50. Two inclusion criteria were used: nationality (i.e., Ireland, the United Kingdom, and the United States) and mother tongue (i.e., English). Employment status of participants was: 48.4% fully employed, 14.9% employed part-time, 11.0% self-employed, 8.4% enrolled as students, and 1.1% unemployed.

Participants were excluded based on failed attention or manipulation checks (explained in more detail below). Analyses pertaining to Hypotheses 1 to 3 were conducted on a final sample of 709 participants and analyses pertaining to Hypothesis 4 were calculated on a final sample of 406 participants. Demographic characteristics of the final samples are reported in Appendix A (Table A.1).

2.2. Procedure

After giving informed consent, participants reported, in random order, the extent to which they endorsed new technologies, such as algorithms (*technology endorsement*), used new technologies (*technology use*), and perceived algorithms as sufficiently developed to carry out tasks with and without human input in either recruitment or finance (*perceived technology maturity*). Participants then read a scenario in which John (the fictional victim of the error) was erroneously rejected for a job or a mortgage loan (i.e., domain of use: recruitment vs. finance) leading to a negative outcome that differed in how severely it affected John (i.e., severity: low vs. high). Participants were informed that the error was made by a decision maker who was either a human or an algorithm.

Participants who read about the erring algorithm additionally read information about the algorithm performance relative to human performance (algorithm maturity: no information about performance vs. same performance as human vs. better than human performance; see Appendix B for a presentation of each manipulated element of the

scenario).

We used a between-subject, not fully crossed (varying based on the decision maker condition) design: 2 (domain of use: recruitment vs. finance) by 2 (severity: low vs. high) by 2 (decision maker: algorithm vs. human). Prior to the study, the authors carried out a pre-test designed to assess participants' perception of the severity and of the credibility of the scenarios. The results (see Appendix C) indicated that the manipulation of severity was successful in creating significant differences in terms of perceived severity in recruitment and in finance and that there was no difference in credibility among the low and high severity scenarios in both domains of use (recruitment and finance).

The manipulation of the information about algorithmic performance (maturity) only concerned the algorithmic decision maker conditions. These had the following design: 2 (domain of use: recruitment vs. finance) by 2 (severity: low vs. high) by 3 (maturity: no information, same as human, better than human). Fig. 1 summarizes the study design. Participants were randomly assigned to read one of the 16 scenarios.

After reading the scenario, participants reported their reactions to the error made by the decision maker. First, participants indicated to what extent they found the error acceptable (i.e., *acceptance*) and then they reported the extent to which the scenario made them feel sad, angry, fearful, and disgusted (i.e., *negative emotions*). They also reported the extent to which they *blamed* and *forgave* the decision maker and they reported their *behavioral intentions* toward the decision maker (e.g., improve/train). The wording of the items was adapted to the decision maker and the domain of use. In the human decision maker condition, we asked about the "HR recruiter" or the "loan underwriter", depending on the domain of use. In the algorithmic decision maker condition, we asked about the "algorithm used for recruitment" or the "algorithm used for loan underwriting", depending on the domain of use. Participants then indicated the extent to which they held the decision maker accountable for the error (*accountability*) and the likelihood of the decision maker making such mistakes (*error proneness*). After responding to manipulation and attention check items, participants provided demographic information (gender, age, and education level) and information related to their professional experience in recruitment and in finance.

2.3. Measures

2.3.1. Perception of technology

Technology Endorsement. Technology endorsement was measured with the technological readiness index composed of 16 items, including eight reverse-scored items (A. Parasuraman & Colby, 2015). A sample item is: "People are too dependent on technology to do things for them" (reverse-scored). All items were rated on a 5-point Likert-type scale (1 = *strongly disagree* to 5 = *strongly agree*). Items were averaged ($M = 3.34$, $SD = 0.56$, $\alpha = 0.85$) and a higher value indicates a stronger tendency to endorse new technologies.

Technology Use. Technology use was measured with three items developed by the authors and rated on a 5-point Likert-type scale (1 = *strongly disagree* to 5 = *strongly agree*). A sample item is: "New technology

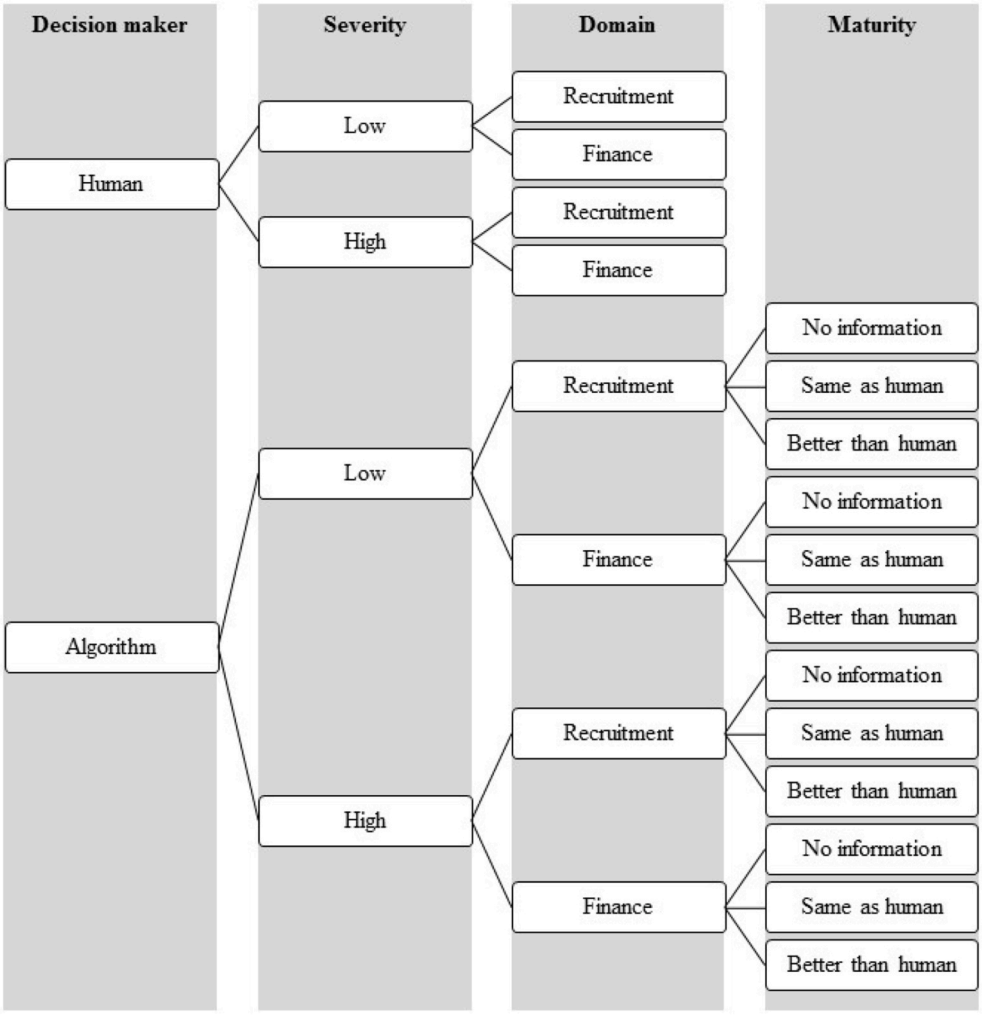


Fig. 1. Experimental design.

is something I often use". Items were averaged ($M = 3.84$, $SD = 0.70$, $\alpha = 0.77$) and a higher value indicates that participants are more familiar with new technologies.

Perceived Technology Maturity. Perceived technology maturity was measured with four items developed by the authors to assess perception of technology maturity in two domains of use (recruitment vs. finance) for two levels of automation (perform a task with vs. without human input). Sample items are "HR recruitment algorithms are sufficiently developed to recruit candidates for a job, without the input of a human recruiter" and "Algorithms used for mortgage loan underwriting in banks are sufficiently developed to assist human underwriters in underwriting mortgage loans for clients". All items were rated on a 5-point Likert-type scale, 1 = *strongly disagree* to 5 = *strongly agree* (without human input, in recruitment: $M = 2.37$, $SD = 1.03$, in finance: $M = 2.79$, $SD = 0.99$; with human input, in recruitment: $M = 3.23$, $SD = 1.03$, in finance: $M = 3.39$, $SD = 0.91$).

Paired *t*-tests showed that participants perceived algorithms as more mature to perform tasks in finance than in recruitment, for both levels of automation, with human input: $t(789) = 4.78$, $p < .001$, and without human input: $t(789) = 12.57$, $p < .001$.

2.3.2. Reactions

The matrix of correlation coefficients concerning all dependent variables is presented in Appendix A (Table A. 2).

Acceptance. Acceptance of the error was measured with eight items (three reverse-scored items) developed by the authors. Participants rated each item using a 5-point scale (1 = *strongly disagree* to 5 = *strongly agree*). A sample item is: "I think that it is inadmissible that such mistakes happen" (reverse scored). Acceptance was computed by averaging the eight items ($M = 2.64$, $SD = 0.62$, $\alpha = 0.78$). A higher value indicates higher acceptance of the error.

Negative Emotions. Negative emotions toward the consequences of the error were measured with 12 items developed by the authors. A sample item is: "The consequences of the mistake described in the scenario make you feel … [adjective]". Each emotion was assessed using three adjectives: "sad/distressed/heartbroken" for sadness, "angry/outraged/furious" for anger, "fearful/frightened/scared" for fear, and "appalled/disgusted/shocked" for disgust. Participants rated each adjective using a 5-point Likert-type scale (1 = *strongly disagree* to 5 = *strongly agree*). The negative emotions score was computed by averaging the 12 adjectives ($M = 3.18$, $SD = 0.81$, $\alpha = 0.93$). A higher value indicates that participants reported feeling more negatively toward the consequences of the error.

Blame. To what extent participants blamed the decision maker for the error was measured with seven items (two reverse-scored items) developed by the authors, including two items based on the Victimization Subscale (Wade, 1989). A sample item is: "I blame the decision maker for making the mistake". Participants rated each item on a 5-point Likert-type scale (1 = *strongly disagree* to 5 = *strongly agree*). Blame was computed by averaging the items ($M = 3.45$, $SD = 0.73$, $\alpha = 0.82$). A higher value indicates a higher tendency to blame the decision maker.

Forgiveness. To what extent participants forgave the decision maker for the error was measured with two items (developed by the authors). A sample item is: "The decision maker should be forgiven for the mistake". Participants rated both items on a 5-point Likert-type scale (1 = *strongly disagree* to 5 = *strongly agree*). Forgiveness was computed by averaging both items ($M = 2.56$, $SD = 0.87$, $r = 0.59$). A higher value indicates a higher tendency to forgive the decision maker.

Accountability. To what extent participants held the decision maker accountable for the error was measured with a four-item scale (one reverse-scored item) developed by the authors. A sample item is: "The decision maker is fully responsible for the mistake." Participants rated each item on a 5-point Likert-type scale (1 = *strongly disagree* to 5 = *strongly agree*). Accountability was computed by averaging the items ($M = 3.32$, $SD = 0.78$, $\alpha = 0.72$). A higher value indicates a higher tendency to perceive the decision maker as accountable for the error.

Behavioral Intentions. Behavioral intentions toward the decision maker after learning about the error and its consequences were measured with four items (developed by the authors). Participants reported the extent to which they would (1) improve/train ($M = 4.50$, $SD = 0.68$), (2) stop using/fire ($M = 3.17$, $SD = 1.12$), (3) do nothing ($M = 1.82$, $SD = 0.84$), and (4) keep using the company that uses/employs the decision maker ($M = 2.84$, $SD = 1.03$). A sample item is: "The decision maker should be improved (algorithm)/trained (human)." Participants rated each item on a 5-point Likert-type scale (1 = *strongly disagree* to 5 = *strongly agree*). A higher value indicates a stronger preference to act according to the specific behavioral intention.

Error proneness. We measured how likely the participants thought the decision maker made on average such an error using a single item (developed by the authors): "Thinking of all the decisions the decision maker takes, in how many cases (in percentage of all the decisions taken), the decision maker makes such a mistake?" Participants reported the likelihood of the decision maker making mistakes by using a slider, 0% = *mistakes occur in none of the decisions taken by the decision maker* to 100% = *mistakes occur in all the decisions taken by the decision maker* ($M = 30.90$, $SD = 25.60$).

Given the previous results pertaining to perception of technology maturity showing that participants perceived algorithms used in finance as more mature than algorithms used in recruitment, we tested whether participants judged errors in recruitment as more likely to occur than errors in finance. Participants judged the errors in recruitment as significantly more likely ($M = 34.01$, $SD = 25.52$) than errors in finance ($M = 27.80$, $SD = 25.33$), $t(878) = 3.63$, $p < .001$.

2.3.3. Demographic Characteristics

We collected data on gender, age, education level, and past experience in the domain of finance and recruitment (see Table A1). Experience was assessed with one item per domain (developed by the authors): "Do you have previous work experience in hiring and recruitment/in the loan business?". The items were rated on a 5-point scale, 1 = *no experience at all* to 5 = *very strong experience* (in recruitment: $M = 2.01$, $SD = 1.18$; in finance: $M = 1.23$, $SD = 0.68$).

2.3.4. Attention checks

To identify careless respondents, three attention check questions were included throughout the questionnaire. A sample item is: "This is an attention check, please press neither agree nor disagree." Correct answers (e.g., selecting "neither agree nor disagree" when asked to do so) were coded 1 and incorrect answers were coded 0.

2.3.5. Manipulation checks

Six manipulation check items were included in the online questionnaire. They were designed to assess whether the experimental manipulations had the intended effect on the participants.[1] The two items for the decision maker were: "In the scenario, the mistake was made by a human/by an algorithm." For the domain of use, the two items were: "Did the scenario involve a mortgage loan/recruitment?" The two items for algorithm maturity were: "The scenario provided me with information that the algorithm typically performs better than humans/equally well as humans on the task." Participants rated these six items using a binary scale ("yes" or "no").

2.4. Data exclusion

To ensure data quality, we first excluded participants who failed two out of three attention checks. Second, we excluded participants based on their answers to the manipulation checks in that we excluded data from participants who failed at least one out of two manipulation checks related to the decision maker, or at least one out of two manipulation checks related to the domain of use (a total of 171 participants were excluded). Concerning algorithm maturity (only participants from the algorithmic decision maker condition), we excluded participants who failed at least one out of two manipulation checks related to algorithm maturity (a total of 258 participants were excluded).

3. Results

3.1. Effects of decision maker on reactions

We calculated 2 (decision maker: human vs. algorithm) by 2 (domain of use: recruitment vs. finance) by 2 (outcome severity: low vs. high) ANOVAs[2] for each of the dependent variables (i.e., acceptance, negative emotions, blame, forgiveness, accountability, and each behavioral intention; see Table 1). To test our hypotheses, we focus on reporting the results concerning the decision maker (main effects of decision maker and interaction effects involving decision maker).

Results (Table 1) showed that the same error with the same consequences was judged as significantly less acceptable when made by an algorithm ($M = 2.59$, $SD = 0.61$) than when made by a human ($M = 2.81$, $SD = 0.64$). Moreover, participants reported significantly more negative feelings with respect to the outcomes of the error, when the error was committed by an algorithm ($M = 3.21$, $SD = 0.80$) compared to when the error was committed by a human ($M = 3.04$, $SD = 0.85$). These results support Hypothesis 1, stating that when learning about an error made by an algorithm, participants' *gut reactions* are harsher (i.e., less acceptance and more negative emotions) than when learning about the same error made by a human.

Participants blamed the human significantly more ($M = 3.81$, $SD = 0.62$) than they blamed the algorithm ($M = 3.38$, $SD = 0.74$) for the same error with the same consequences (see Table 1). Additionally, participants forgave the erring human significantly more ($M = 2.75$, $SD = 0.74$) than the erring algorithm ($M = 2.47$, $SD = 0.87$). Finally, participants considered the erring human as significantly more accountable for the error ($M = 3.85$, $SD = 0.60$) than the erring algorithm ($M = 3.23$, $SD = 0.76$). These results support Hypothesis 2, stating that when learning about an error made by an algorithm, participants' *justice cognitions* are weaker (i.e., less blame, less forgiveness, less accountability) than when learning about the same error made by a human.

Participants thought that the erring algorithm should be improved ($M = 4.64$, $SD = 0.58$) significantly more so than they thought the erring human should be trained ($M = 4.17$, $SD = 0.80$) for the same error with the same consequences (see Table 1). Participants thought that one should stop using the erring algorithm ($M = 3.38$, $SD = 1.08$) significantly more so than they thought that one should fire the erring collaborator ($M = 2.44$, $SD = 0.92$). Participants thought that nothing can be done when the human made the error ($M = 2.12$, $SD = 0.83$) significantly more so than when the algorithm made the error ($M = 1.70$, $SD = 0.81$). There was no significant difference with respect to "keep using the company", regardless of whether the error was made by an algorithm ($M = 2.80$, $SD = 1.02$) or a human ($M = 2.83$, $SD = 1.02$). These results support Hypothesis 3, stating that when learning about an error made by an algorithm, participants' *behavioral intentions* are to act (i.e., improve/train, stop using/fire, nothing can be done, or keep using), more so than when learning about the same error committed by a human, except for keep using the company, which did not show a difference according to the decision maker.

To find out whether the severity of the outcome or the domain of use in which the error occurred affected the results differently for the algorithmic or the human decision maker, we looked at the interaction effects involving the decision maker. Table 1 shows that there was only one significant interaction effect. One aspect of the behavioral intentions, namely whether humans or algorithms should be improved, depended on the severity of the outcome. The simple main effect analysis showed that participants thought that humans should be trained when severity was low, $M = 4.30$, $SE = 0.07$, as compared to when severity was high, $M = 4.03$, $SE = 0.07$, $F(1,701) = 6.73$, $p = .010$. However, when the decision maker was an algorithm, severity did not affect the extent to which participants thought the algorithm should be improved, $F(1,701) = 0.15$, $p = .699$.

[1] The manipulations were successful. All results were in the expected direction and were significant at $p < .001$. Participants who read the human decision maker scenario indicated more often that the decision maker was a human than participants who read the algorithm decision maker scenario, χ^2 (1, $N = 875$) = 463.52, $p < .001$. Participants who read the algorithm decision maker scenario indicated more often that the decision maker was an algorithm than participants who read the human decision maker scenario, χ^2 (1, $N = 876$) = 492.01, $p < .001$.Participants who read the finance scenario indicated more often that the scenario was about a mortgage loan than participants who read the recruitment scenario, χ^2 (1, $N = 876$) = 680.17, $p < .001$. Participants who read the recruitment scenario indicated more often that the scenario was about recruitment than participants who read the finance scenario, χ^2 (1, $N = 875$) = 789.72, $p < .001$.Participants who read the scenario specifying that algorithms perform better than humans indicated more often that algorithms outperform humans than participants who read the scenario indicating that algorithms perform equally well as humans, χ^2 (1, $N = 436$) = 122.61, $p < .001$. Participants who read the scenario indicating that algorithms perform equally well as humans more often said that both perform equally well compared to participants who read the scenario indicating that algorithms outperform humans, χ^2 (1, $N = 437$) = 98.08, $p < .001$.

[2] We additionally performed ANCOVA analyses to test whether controlling for perception of technology (i.e., technology endorsement, technology use and perceived technology maturity relative to the domain of use), past experience in recruitment and in finance, and socio-demographics (i.e., gender and age) affected our results. The ANCOVA results showed that the inclusion of these control variables did not change our results concerning decision maker, except that one previously significant effect became marginal when using technology endorsement and technology use as control variables. This concerned the main effect of decision maker on negative emotions.

[3] We additionally performed ANCOVA analyses to test whether controlling for perception of technology (i.e., technology endorsement, technology use and perceived technology maturity relative to the domain of use), past experience in recruitment and in finance, and socio-demographics (i.e., gender and age) affected our results. Our results concerning maturity remained stable.

Table 1
Analyses of variance testing for the effect of decision maker, severity, and domain of use, on gut reactions, justice cognition, and behavioral intentions.

	Gut Reactions					
	Acceptance			Negative Emotions		
	$F(1,701)$	p	η^2	$F(1,701)$	p	η^2
DM	15.16	.000	.02	4.85	.028	.01
Severity	4.11	.043	.01	10.41	.001	.01
Domain of use	7.33	.007	.01	1.82	.178	.00
DM*Severity	0.40	.525	.00	0.22	.636	.00
DM*Domain of use	0.14	.704	.00	0.12	.733	.00
Severity*Domain of use	2.61	.106	.00	3.61	.058	.01
DM*Severity*Domain of use	0.66	.417	.00	0.27	.603	.00

	Justice Cognitions								
	Blame			Forgiveness			Accountability		
	$F(1,701)$	p	η^2	$F(1,701)$	p	η^2	$F(1,701)$	p	η^2
DM	43.78	.000	.06	13.39	.000	.02	89.19	.000	.11
Severity	0.00	.993	.00	0.29	.592	.00	0.56	.456	.00
Domain of use	0.50	.481	.00	1.55	.213	.00	0.01	.939	.00
DM*Severity	0.32	.573	.00	0.02	.900	.00	0.00	.982	.00
DM*Domain of use	1.29	.257	.00	0.02	.901	.00	2.16	.142	.00
Severity*Domain of use	0.36	.551	.00	0.23	.631	.00	1.45	.229	.00
DM*Severity*Domain of use	2.90	.089	.00	0.00	.994	.00	0.16	.689	.00

	Behavioral Intentions											
	Improve			Fire			Nothing			Keep using		
	$F(1,701)$	p	η^2	$F(1,701)$	p	η^2	$F(1,701)$	p	η^2	$F(1,701)$	p	η^2
DM	69.77	.000	.09	96.32	.000	.12	31.97	.000	.04	0.03	.856	.00
Severity	4.47	.035	.01	6.50	.011	.01	0.44	.505	.00	5.44	.020	.01
Domain of use	1.10	.295	.00	0.05	.823	.00	2.97	.085	.00	0.12	.724	.00
DM*Severity	6.13	.014	.01	2.12	.145	.00	0.03	.852	.00	2.46	.117	.00
DM*Domain of use	0.65	.422	.00	0.38	.538	.00	1.69	.195	.00	2.59	.108	.00
Severity*Domain of use	1.08	.298	.00	1.61	.204	.00	3.89	.049	.01	1.54	.214	.00
DM*Severity*Domain of use	0.88	.349	.00	0.24	.624	.00	1.83	.176	.00	0.52	.472	.00

Note. $n = 709$. DM = Decision maker.

3.2. Effects of maturity on reactions

To find out how algorithm maturity affects the results for algorithmic decision makers, we calculated 2 (domain of use: recruitment vs. finance) by 2 (outcome severity: low vs. high) by 3 (maturity: no information, same as human, better than human) ANOVAs[3] for each of the dependent variables (i.e., acceptance, negative emotions, blame, forgiveness, accountability, and behavioral intentions). If the results of the ANOVAs concerning the algorithm maturity factor were significant, we performed pairwise contrast analyses to compare the "same as human" and "better than human" conditions.

Results of the ANOVAs were significant only for *gut reactions* in that a main effect of maturity was observed for acceptance and negative emotions (Table 2). The contrast analyses between the two conditions of "same as human" and "better than human" yielded no significant difference in terms of acceptance (Fig. 2), $p = .522$, and with respect to negative emotions (Fig. 3), $p = .895$. Thus, hypothesis 4, stating that *gut reactions* are harsher (i.e., less acceptance and more negative emotions) when being informed that algorithms perform better than humans (compared to being informed that algorithms perform at the same level as humans), was not confirmed.

4. Discussion

We set out to test different types of reactions (i.e., gut reactions in the form of acceptance and negative emotions; justice cognitions comprising blame, forgiveness, and perceived accountability, as well as behavioral intentions) toward an erring algorithmic decision maker compared to an erring human decision maker. We also investigate whether the maturity of the algorithm (i.e., how well it performs compared to a human for the same task), the severity of the error, and the domain of use (i.e., recruitment vs. finance) affect those reactions.

Our results confirm that to err is human, not algorithmic. On a rather implicit, instinctive level, participants reported lower acceptance and more negative emotions toward the erring algorithm than toward the erring human, confirming Hypothesis 1. On a more cognitively elaborate level (justice cognitions), our results also confirm the non-human nature of how algorithms are perceived; they were subject to less blame, less forgiveness, and less perceived accountability than the erring human, confirming Hypothesis 2. On the behavioral intention level, our results showed that people intended to act (to improve/train and to stop using/fire) when confronted with an erring algorithm more so than when confronted with an erring human, confirming Hypothesis 3. These results suggest that we hold algorithms to higher performance standards (the perfect algorithm vs. the imperfect human, Madhavan & Wiegmann, 2007) and are thus less accepting of algorithmic errors and want to act upon an erring algorithm by either improving it or simply by stopping using it. They also show that algorithms are less subject to reactions usually reserved to humans (i.e., blame, forgiveness, accountability).

Although research on algorithm aversion has shown that people are less willing to trust and use erring algorithms compared to erring humans, our research contributes to this literature in that we look at a broader array of reactions according to the third-party reactions to mistreatment model (O'Reilly & Aquino, 2011). We show that people do not simply react negatively to erring algorithms (Dietvorst et al., 2014, 2018; Dzindolet et al., 2003; Prahl & Van Swol, 2017). For instance, people blame the erring algorithm *less* than an erring human, showing that justice cognitions are reserved for humans, as already suggested by previous authors (see Malle et al., 2014; Pizarro, 2014). Therefore, an erring algorithm does not just elicit a generalized negative reaction on all levels (instinctive, cognitive-moral, and behavioral), but a rather

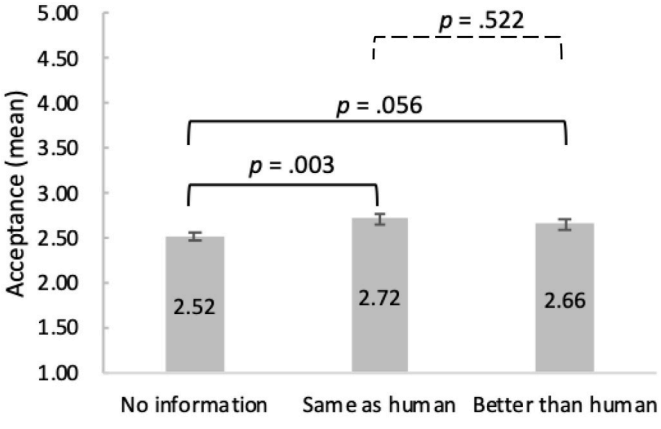
Table 2
Analyses of Variance Testing for the Effect of Maturity, Severity, and Domain of Use, on Gut Reactions, Justice Cognitions, and Behavioral Intentions, for Algorithmic Decision Makers only.

	Gut Reactions					
	Acceptance			Negative Emotions		
	F	p	η^2	F	p	η^2
Maturity	4.77[a]	.009	.02	4.55[a]	.011	.02
Severity	4.30[b]	.039	.01	5.03[b]	.025	.01
Domain of use	14.35[b]	.000	.04	6.37[b]	.012	.02
Maturity *Severity	1.27[a]	.281	.01	1.60[a]	.203	.01
Maturity *Domain of use	0.15[a]	.860	.00	0.25[a]	.776	.00
Severity*Domain of use	1.53[b]	.218	.00	1.53[b]	.216	.00
Maturity *Severity*Domain of use	0.78[a]	.458	.00	2.10[a]	.124	.01

	Justice Cognitions								
	Blame			Forgiveness			Accountability		
	F	p	η^2	F	p	η^2	F	p	η^2
Maturity	2.11[a]	.122	.01	0.80[a]	.452	.00	1.35[a]	.262	.01
Severity	0.11[b]	.738	.00	0.56[b]	.456	.00	1.00[b]	.318	.00
Domain of use	5.84[b]	.016	.02	2.77[b]	.097	.01	3.93[b]	.048	.01
Maturity *Severity	1.46[a]	.233	.01	0.44[a]	.647	.00	0.76[a]	.470	.00
Maturity *Domain of use	0.42[a]	.656	.00	0.52[a]	.594	.00	0.04[a]	.959	.00
Severity*Domain of use	3.81[b]	.052	.01	0.26[b]	.608	.00	0.84[b]	.359	.00
Maturity *Severity*Domain of use	0.04[a]	.959	.00	0.41[a]	.667	.00	0.13[a]	.882	.00

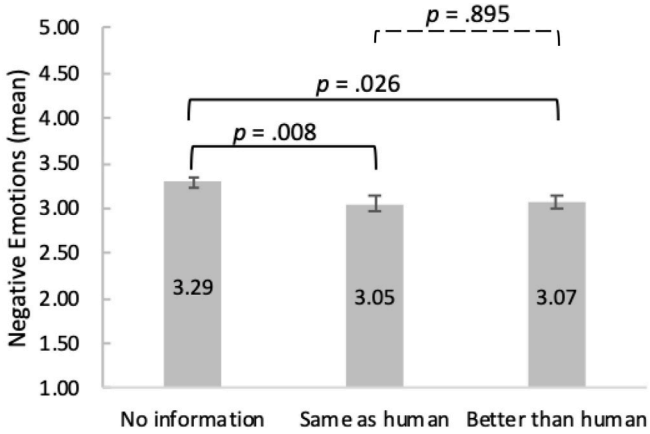
	Behavioral Intentions											
	Improve			Fire			Nothing			Keep using		
	F	p	η^2	F	p	η^2	F	p	η^2	F	p	η^2
Maturity	0.19[a]	.825	.00	2.54[a]	.080	.01	0.71[a]	.492	.00	0.71[a]	.492	.00
Severity	0.05[b]	.827	.00	1.62[b]	.204	.00	0.01[b]	.923	.00	0.33[b]	.568	.00
Domain of use	3.66[b]	.056	.01	0.19[b]	.662	.00	5.16[b]	.024	.01	0.01[b]	.916	.00
Maturity *Severity	0.06[a]	.941	.00	0.37[a]	.691	.00	0.81[a]	.445	.00	1.02[a]	.363	.01
Maturity *Domain of use	0.31[a]	.736	.00	0.09[a]	.916	.00	1.59[a]	.205	.01	0.54[a]	.582	.00
Severity*Domain of use	0.26[b]	.611	.00	0.01[b]	.911	.00	0.90[b]	.344	.00	0.03[b]	.867	.00
Maturity *Severity*Domain of use	0.60[a]	.548	.00	0.21[a]	.810	.00	1.65[a]	.194	.01	0.88[a]	.414	.00

Note. $n = 406$.
[a] $F(2,394)$.
[b] $F(1,394)$.



Note. Error bars show standard errors.

Fig. 2. Effect of Information about Algorithm Performance (Algorithm Maturity) on Acceptance. *Note*. Error bars show standard errors.



Note. Error bars show standard errors.

Fig. 3. Effect of Information about Algorithm Performance (Algorithm Maturity) on Negative Emotions. *Note*. Error bars show standard errors.

differentiated reaction pattern that is more in line with being perceived as non-human: negative gut reactions, no human-typical moral or justice cognitions, and a utilitarian, functional approach to behavioral intentions (i.e., do something about it by trying to fix it or stop using it).

It has to be noted that our results concerning negative emotions differ from Prahl and Van Swol's results (2017). Alongside studying the effect of receiving bad computer-based (vs. human-based) advice, these authors also tested whether emotional reactions to the ill-advising algorithm and human differed. After the task, participants reported their positive (i.e., appreciative, happy, grateful, thankful, satisfied, and glad)

and negative emotions (i.e., mad, frustrated, annoyed, and irritated) when receiving advice. Their results show no difference concerning emotional reactions (positive and negative) toward algorithms or humans. This difference to our results might originate in differences in the methods used. For instance, we assessed a variety of negative emotions (i.e., sadness, anger, fear, and disgust) while Prahl and Van Swol focused on only one negative emotion (i.e., anger). It is plausible that differences in emotional reactions due to the decision maker go beyond anger and that the reaction is more of a generalized negative emotional reaction, less driven by anger alone. However, when looking at only the anger-related adjectives in our study (i.e., angry, outraged, furious) our results remained unchanged. Participants reported more anger toward an erring algorithm than toward an erring human. The fact that the erring algorithm in Prahl and Van Swol's study did not directly affect an individual, while in our study the error directly affected John (see Appendix B – severity), might explain the different results. It is plausible that our design triggered harsher moral anger because the error affected a human (see O'Reilly & Aquino, 2011).

When testing whether information about algorithm performance (algorithm maturity) affected the reactions toward erring algorithms, our results show no effect, and hence do not confirm Hypothesis 4. Therefore, even if we provide information about how performant the algorithm is, and even when it is common knowledge that on average algorithms outperform humans in a given task, reactions toward erring algorithms are unaffected by this. When learning about an error made by the algorithm, participants seem outraged and are oblivious to information about the typical algorithm performance. They may think that making a mistake means that the algorithm is simply "broken" and they do not care whether it usually performs better than a human. If we expect an algorithm to be perfect, the notion of making less errors than humans might not be relevant–either it is perfect or it is inefficient. Algorithms are perceived as being typically accurate all the time, unlike humans. Our results raise a number of questions about how to communicate algorithm performance and maturity in order for humans to trust them and to continue to use them when they are helpful for decision making even after having made an error. Future work should determine how communication about algorithm performance is linked to algorithm aversion (e.g., Burton et al., 2020; Logg, Minson, & Moore, 2019).

Our results also show that reactions to erring algorithms are immune to the severity of the error and to the domain in which the error occurred (recruitment or finance). Algorithms might simply be "stereotyped" as non-human–especially when they make mistakes–and be immune to a more differentiated view of them.

The same undifferentiated result emerges with respect to the domain of use. This is particularly interesting because we showed in this study that perceived technology maturity differs according to domain: People perceive algorithms to be less mature to perform recruitment tasks than to perform financial tasks with or without human input (consistent with Castelo and colleagues' research, 2019). Therefore, the fact that our participants perceived algorithms in finance to be more mature than algorithms in recruitment did not affect their reactions.

Although research in algorithm aversion has often used consequential measures of participant behavior such that their pay depended on an algorithm-based or a human-based estimation (see similar examples in Dietvorst et al., 2014, and Prahl & Van Swol, 2017), we opted for using self-reported measures. This choice was driven by us aiming at measuring intuitive, cognitive, and behavioral reactions with the same method, and thus not to introduce differences with regard to the type of assessment. Moreover, gut reactions and moral judgments are typically assessed via self-report, which is why we used self-report in our study.

Similarly, one might criticize the use of vignettes over the use of a more experiential approach as done by others (see Dietvorst et al., 2018, 2014; Dzindolet et al., 2003; Prahl & Van Swol, 2017). Given that the aim of our study was to test several types of boundary conditions, the use of vignettes was not only the most efficient approach, it ensured high experimental control. Additionally, research suggests that even without experiencing algorithm errors oneself, learning about others' (un) fortunate encounters with algorithms affects their adoption of said algorithms (see Alexander et al., 2018; Shariff et al., 2017). This is consistent with the deontic justice literature according to which people might seek to wrong what they consider an unfair treatment (react to someone else mistreatment) even if it does not have an effect of them (see, for example, Skarlicki, Ellard, & Kelln, 1998; Turillo, Folger, Lavelle, Umphress, & Gee, 2002).

In the vignettes, we used a third-person rather than a first-person perspective. The reason for this choice was that participants most likely differ in their experience as a job applicant or as a loan applicant which will affect their perception and reaction to the vignettes if they are posed in the first-person perspective much more than when they read about it in the third-person perspective. Observing (third-person perspective) an error or mistreatment might however lead to less intense reactions than experiencing (first-person perspective) it (Jones, 2011; Kray & Lind, 2002; Lind, Kray, & Thompson, 1998). Therefore, the result of seeing the algorithms as non-human and reacting to them in such a way might be an underestimation of the true effect.

Finally, we did not assess potential mechanisms underlying the observed effects. For instance, future research could establish the role of perfect automation schemas in explaining reactions to erring algorithms. An avenue of research would be to include an assessment of such beliefs beyond assessing the perception of technology. Doing so would enable researchers to determine the role of cognitive schemas in explaining negative reactions to erring algorithms.

To conclude, we show that intuitive, cognitive, and behavioral reactions to erring algorithms indicate that algorithms are perceived and reacted upon as "non-human"; they are not allowed to err and we are less lenient in our judgment toward them when they do. Moreover, those reactions are very robust in that we do not take into account how well such algorithms perform in general nor how bad the mistake was when they erred, nor do we differentiate much according to the domain of use.

Credit author statement

Laetitia A. Renier: Conceptualization, Methodology, Resources, Investigation, Data curation, Formal analysis, Writing - Original Draft. Marianne Schmid Mast: Funding acquisition, Conceptualization, Methodology, Resources, Investigation, Writing - Review & Editing. Anely Bekbergenova: Conceptualization, Methodology, Data curation, Writing - Review & Editing.

Statement of ethical compliance

All procedures performed in the following studies are in accordance with the ethical standards of institutional research committees and with the 1964 Helsinki declaration and its later amendments for treatment of human participants. The authors declare no competing interests in the conduct and publication of this research.

Data availability

The data, syntaxes, and codebook are available on OSF repository: https://osf.io/jn3gk/?view_only=1b6c928aabd04ef8a8b5ef339683ec3b.

Declaration of competing interest

None.

Acknowledgments

This research was supported by a grant from the Swiss National Science Foundation – SNF (grant reference: SINERGIA CRSII5_183564).

Appendix A

A.1 Demographic characteristics of the final samples

Table A.1
Demographic Characteristics of Study Participants.

	Characteristics	Human and Algorithmic Decision Maker		Algorithmic Decision Maker	
		n = 709	%	*n* = 406	%
Gender	Women	364	51.30	202	49.80
	Men	342	48.20	202	49.80
	Unknown	3	0.40	2	0.50
Age	*M*	36.57		35.91	
	SD	12.05		11.40	
	Min-Max	18–81		18–75	
Education level	Some high school, no diploma	20	2.80	8	2.00
	High school degree or equivalent	216	30.50	106	26.10
	Apprenticeship	28	3.90	10	2.50
	Bachelor's degree	308	43.40	200	49.30
	Master's degree	88	12.40	50	12.30
	Doctorate (e.g. PhD)	15	2.10	8	2.00
	Other	21	3.00	14	3.40
Employment status	Student	57	8.00	35	8.60
	Self-employed	83	11.70	53	13.10
	Unemployed	69	9.70	31	7.60
	Employed full-time	342	48.20	201	49.50
	Employed part-time	104	14.70	57	14.00
	Other	48	6.80	23	5.70
Perception of Technology	Technology endorsement				
	M	3.34		3.38	
	SD	0.57		0.55	
	Technology use				
	M	3.84		3.88	
	SD	0.70		0.66	

A.2 Matrix of correlation coefficients pertaining to the dependent variables

Table A. 2
Pearson Correlations Concerning the Dependent Variables.

		1		2		3		4		5		6		7		8		9	
1	Acceptance																		
2	Negative emotions	-.58	***																
3	Blame	-.29	***	.27	***														
4	Forgiveness	.40	***	-.27	***	-.43	***												
5	Perceived Accountability	-.12	**	.18	***	.70	***	-.32	***										
6	Probability	-.14	***	.21	***	.19	***	-.10	**	.15	***								
7	Improve/train	-.18	***	.03		-.01		-.15	***	-.09	*	-.09	*						
8	Stop using/fire	-.35	***	.31	***	.19	***	-.36	***	.11	**	.23	***	.08	*				
9	Nothing can be done	.25	***	-.07	*	-.04		.26	***	.06		.11	**	-.31	***	-.19	***		
10	Keep using	.40	***	-.31	***	-.27	***	.33	***	-.21	***	-.21	***	.01		-.43	***	.21	***

Note. n = 709.
*p < .05. **p < .01. ***p < .001.

Appendix B

Presentation of the Manipulated Elements of the Scenarios

Manipulation		
Domain of use	**Recruitment** John, an individual, who was struggling financially, applied for an interesting job opportunity. John had the necessary job skills, rich experience in the field, and a relevant professional background. Nevertheless, he received a rejection letter from the company.	**Finance** John, an individual, who was struggling financially, applied for a mortgage loan renewal for his house. John had a strong application, and was compliant to all the requirements to receive an approval for the mortgage request. Nevertheless, his mortgage loan request was declined by the local bank.
Severity	**Low** Following this job rejection (*finance*: loan rejection), John continued to struggle financially. He had to cut his food and leisure expenses for a couple of months to keep his house.	**High** Following this job rejection (*finance*: loan rejection), John entered a period of great financial hardship, got evicted from his house, and had to stay with a friend for a couple of months.

(*continued on next page*)

(*continued*)

Manipulation			
Decision Maker	**Human**		**Algorithm**
	It was established that the rejection was a hiring mistake made entirely by the HR recruiter of the company (*finance*: loan underwriter in the bank). John should have been hired (*finance*: received the loan).		It was established that the rejection was a hiring mistake made entirely by the computer algorithm used for recruitment (*finance*: for loan underwriting) by the company. John should have been hired (*finance*: received the loan).
Maturity	No information	Same as human Typically such algorithms used for recruitment (*finance*: loan underwriting) perform at the same level and make the same amount of mistakes as human recruiters (*finance*: loan underwriters).	Better than human Typically such algorithms used for recruitment (*finance*: loan underwriting) perform better and make less mistakes than human recruiters (*finance*: loan underwriters).

Appendix C

Results of the Pre-Test

Prior to the study, the authors carried out a pre-test designed to assess participants' perception of the severity and the credibility of the scenarios (stripped from the decision maker and performance elements). The authors carried out a 2 (domain of use: recruitment vs. finance, within-subject variable) by 3 (severity: low vs. high vs. extreme, between-subject variable) mixed design study on a sample of 178 participants.

Concerning perceived severity, the results of the two-way mixed ANOVA showed that there was a significant main effect of domain of use, $F(1,175) = 17.22$, $p < .001$, $\eta_p^2 = 0.09$, and of severity, $F(2,175) = 40.62$, $p < .000$, $\eta_p^2 = 0.32$. The interaction effect of domain of use and severity was not significant $F(2,175) = 0.79$, $p = .458$, $\eta_p^2 = 0.01$. Participants perceived the errors as significantly more severe in Finance ($M = 3.84$, $SD = 0.58$) than in HR ($M = 3.69$, $SD = 0.64$).

For the main effect of severity in HR, $F(2,175) = 33.25$, $p < .001$, pairwise contrast analyses showed that participants perceived the extremely severe outcome as significantly more severe ($M = 4.06$, $SD = 0.46$) than the high severe outcome ($M = 3.79$, $SD = 0.58$), $t(175) = 2.64$, $p = .009$. They perceived the high severe outcome as significantly more severe than the low severe outcome ($M = 3.26$, $SD = 0.60$), $t(175) = 5.30$, $p < .001$.

For the main effect of severity in Finance, $F(2,175) = 32.28$, $p < .001$, pairwise contrasts analyses showed that participants perceived the extremely severe outcome as significantly more severe ($M = 4.15$, $SD = 0.49$) than the high severe outcome ($M = 3.96$, $SD = 0.44$), $t(175) = 2.08$, $p = .039$. They perceived the high severe outcome as significantly more severe than the low severe outcome ($M = 3.45$, $SD = 0.55$), $t(175) = 5.63$, $p < .001$.

Concerning perceived credibility, the results of the two-way mixed ANOVA showed that there was no significant main effect of domain of use, $F(1,175) = 0.151$, $p = .698$, $\eta_p^2 = 0.001$, and a non-significant main effect of Severity, $F(2,175) = 2.886$, $p = .058$, $\eta_p^2 = 0.03$. The interaction effect of domain of use and Severity was also not significant $F(2,175) = 0.79$, $p = .458$, $\eta_p^2 = 0.01$.

References

Alexander, V., Blinder, C., & Zak, P. J. (2018). Why trust an algorithm? Performance, cognition, and neurophysiology. *Computers in Human Behavior, 89*, 279–288. https://doi.org/10.1016/j.chb.2018.07.026

Brooks, D. J., Begum, M., & Yanco, H. A. (2016). Analysis of reactions towards failures and recovery strategies for autonomous robots. In *25th IEEE international symposium on robot and human interactive communication* (pp. 487–492). RO-MAN *2016*. https://doi.org/10.1109/ROMAN.2016.7745162.

Brzezicki, M. A., Bridger, N. E., Kobetić, M. D., Ostrowski, M., Grabowski, W., Gill, S. S., et al. (2020). Artificial intelligence outperforms human students in conducting neurosurgical audits. *Clinical Neurology and Neurosurgery, 192*(February). https://doi.org/10.1016/j.clineuro.2020.105732

Burton, J. W., Stein, M., & Jensen, T. B. (2020). A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making, 33*(2), 220–239. https://doi.org/10.1002/bdm.2155

Campbell, M., Hoane, A. J., & Hsu, F. H. (2002). Deep blue. *Artificial Intelligence, 134* (1–2), 57–83. https://doi.org/10.1016/S0004-3702(01)00129-1

Castelo, N., Bos, M. W., & Lehmann, D. R. (2019). Task-dependent algorithm aversion. *Journal of Marketing Research, 56*(5), 809–825. https://doi.org/10.1177/0022243719851788

Dastin, J. (2018). *Amazon scraps secret AI recruiting tool that showed bias against women - Reuters*. Reuters. Retrieved from https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G.

Derous, E., & De Fruyt, F. (2016). Developments in recruitment and selection management essay. *International Journal of Selection and Assessment, 24*(1), 1. Retrieved from http://www.ukessays.com/essays/management/developments-in-recruitment-and-selection-management-essay.php?cref=1.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2014). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General, 144*(1), 114–126. https://doi.org/10.1037/xge0000033

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2018). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science, 64*(3), 1155–1170. https://doi.org/10.1287/mnsc.2016.2643

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. *International Journal of Human-Computer Studies, 58*, 697–718. https://doi.org/10.1016/S1071-5819(03)00038-7

van Esch, P., Black, J. S., & Ferolie, J. (2019). Marketing AI recruitment: The next phase in job application and selection. *Computers in Human Behavior, 90*(April 2018), 215–222. https://doi.org/10.1016/j.chb.2018.09.009

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors, 3*, 407–434. https://doi.org/10.1177/0018720814547570

Jones, K. S. (2011). *"I'm too good to be bad": The moderating role of honesty-humility in aggressive and prosocial reactions to first and third party unfairness. PhD dissertation.* University of Illinois.

Jussupow, E., Benbasat, I., & Heinzl, A. (2020). Why are we averse towards algorithms? A comprehensive literature review on algorithm aversion. In *ECIS 2020* (pp. 1–16).

Kray, L. J., & Lind, E. A. (2002). The injustices of others: Social reports and the integration of others' experiences in organizational justice judgments. *Organizational Behavior and Human Decision Processes, 89*(1), 906–924. https://doi.org/10.1016/S0749-5978(02)00035-3

Lambrecht, A., & Tucker, C. (2019). Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of stem career ads. *Management Science, 65*(7), 2966–2981. https://doi.org/10.1287/mnsc.2018.3093

Lee, C. S., Nagy, P. G., Weaver, S. J., & Newman-Toker, D. E. (2013). Cognitive and system factors contributing to diagnostic errors in radiology. *American Journal of Roentgenology, 201*(3), 611–617. https://doi.org/10.2214/AJR.12.10375

Lind, E. A., Kray, L., & Thompson, L. (1998). The social construction of injustice: Fairness judgments in response to own and others' unfair treatment by authorities. *Organizational Behavior and Human Decision Processes, 75*(1), 1–22. https://doi.org/10.1006/obhd.1998.2785

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes, 151*, 90–103. https://doi.org/10.1016/j.obhdp.2018.12.005

Lowry, S., & Macpherson, G. (1988). A blot on the profession. *British Medical Journal, 296* (6623), 657–658. https://doi.org/10.1136/bmj.296.6623.657

Madhavan, P., & Wiegmann, D. A. (2007). Similarities and differences between human–human and human–automation trust: An integrative review. *Theoretical Issues in Ergonomics Science, 8*(4), 277–301. https://doi.org/10.1080/14639220500337708

Malle, B. F., Guglielmo, S., & Monroe, A. E. (2014). A theory of blame. *Psychological Inquiry, 25*(2), 147–186. https://doi.org/10.1080/1047840X.2014.877340

Markus, M. L., Dutta, A., Steinfield, C. W., & Wigand, R. T. (2008). The computerization movement in the US home mortgage industry: Automated underwriting from 1980 to 2004. In *Computerization movements and technology diffusion: From mainframes to ubiquitous computing* (pp. 115–144).

Miller, A. P. (2018). Want less-biased decisions? Use algorithms. *Harvard Business Review, 26*. Retrieved from https://hbr.org/2018/07/want-less-biased-decisions-use-algorithms.

Naquin, C. E., & Kurtzberg, T. R. (2004). 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. https://doi.org/10.1016/j.obhdp.2003.12.001

O'Reilly, J., & Aquino, K. (2011). A model of third parties' morally motivated responses to mistreatment in organizations. *Academy of Management Review, 36*(3), 526–543. https://doi.org/10.5465/amr.2009.0311

Parasuraman, A., & Colby, C. L. (2015). An updated and streamlined technology readiness index: TRI 2.0. *Journal of Service Research, 18*(1), 59–74. https://doi.org/10.1177/1094670514539730

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors, 39*(2), 230–253. https://doi.org/10.1518/001872097778543886

Patel, V. L., Shortliffe, E. H., Stefanelli, M., Szolovits, P., Berthold, M. R., Bellazzi, R., et al. (2010). The coming of age of artificial intelligence in medicine. *Artificial Intelligence in Medicine, 46*(1), 5–17. https://doi.org/10.1016/j.artmed.2008.07.017

Phillips, P. J., & O'Toole, A. J. (2014). Comparison of human and computer performance across face recognition experiments. *Image and Vision Computing, 32*(1), 74–85. https://doi.org/10.1016/j.imavis.2013.12.002

Pizarro, D. (2014). Androids, algorithms, and the attribution of blame. *Psychological Inquiry, 25*(2), 234–235. https://doi.org/10.1080/1047840X.2014.904691

Prahl, A., & Van Swol, L. (2017). Understanding algorithm aversion: When is advice from automation discounted? *Journal of Forecasting, 36*, 691–702. https://doi.org/10.1002/for.2464

Promberger, M., & Baron, J. (2006). Do patients trust computers ? *Journal of Behavioral Decision Making, 19*, 455–468. https://doi.org/10.1002/bdm.452

Radford, A., Wu, J., Amodei, D., Amodei, D., Clark, J., Brundage, M., et al. (2019a). *Better language models and their implications*. OpenAI. Retrieved from https://openai.com/blog/better-language-models/.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019b). Language models are unsupervised multitask learners. *OpenAI blog, 1*. Retrieved from http://www.persagen.com/files/misc/radford2019language.pdf.

Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). *Squad: 100,000+ questions for machine comprehension of text*. arXiv preprint arXiv:1606.05250.

Robbennolt, J. K. (2000). Outcome severity and judgments of "responsibility": A meta-analytic review. *Journal of Applied Social Psychology, 30*(12), 2575–2609. https://doi.org/10.1111/j.1559-1816.2000.tb02451.x

Shariff, A., Bonnefon, J. F., & Rahwan, I. (2017). Psychological roadblocks to the adoption of self-driving vehicles. *Nat. Human Behav., 1*(10), 694–696. https://doi.org/10.1038/s41562-017-0202-6

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., et al. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature, 529*(7587), 484–489. https://doi.org/10.1038/nature16961

Skarlicki, D. P., Ellard, J. H., & Kelln, B. R. C. (1998). Third-party perceptions of a layoff: Procedural, derogation, and retributive aspects of justice. *Journal of Applied Psychology, 83*, 119–127. https://doi.org/10.1037/0021-9010.83.1.119

SQuAD. The Standford Question Answering Dataset. https://rajpurkar.github.io/SQuAD-explorer/Accessed October 2020.

Straka, J. W. (2000). A shift in the mortgage landscape: The 1990s move to automated credit evaluations. *Journal of Housing Research, 11*(2), 207–232. Retrieved from https://www.jstor.org/stable/24833780.

Tschandl, P., Codella, N., Akay, B. N., Argenziano, G., Braun, R. P., Cabo, H., et al. (2019). Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: An open, web-based, international, diagnostic study. *The Lancet Oncology, 20*(7), 938–947. https://doi.org/10.1016/S1470-2045(19)30333-X

Turillo, C. J., Folger, R., Lavelle, J. J., Umphress, E. E., & Gee, J. O. (2002). Is virtue its own reward? Self-sacrificial decisions for the sake of fairness. *Organizational Behavior and Human Decision Processes, 89*(1), 839–865.

Upadhyay, A. K., & Khandelwal, K. (2018). Applying artificial intelligence: Implications for recruitment. *Strategic HR Review, 17*(5), 255–258. https://doi.org/10.1108/shr-07-2018-0051

Vinyals, O., Babuschkin, I., Czarnecki, W. M., Mathieu, M., Dudzik, A., Chung, J., et al. (2019). Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature, 575*(7782), 350–354. https://doi.org/10.1038/s41586-019-1724-z

Wade, S. H. (1989). *The development of a scale to measure forgiveness*. CA: Pasadena.