Received: 4 February 2022 | Accepted: 13 June 2022

DOI: 10.1002/jcpy.1313

RESEARCH ARTICLE

# How to overcome algorithm aversion: Learning from mistakes

Taly Reich[1] | Alex Kaju[2] | Sam J. Maglio[3]

[1]Yale School of Management, Yale University, New Haven, Connecticut, USA

[2]HEC Montréal, Montréal, Canada

[3]University of Toronto Scarborough & Rotman School of Management, Toronto, Canada

**Correspondence**
Taly Reich, Yale School of Management, Yale University, 165 Whitney Ave., New Haven, Connecticut 06511, USA.
Email: taly.reich@yale.edu

**Abstract**

When consumers avoid taking algorithmic advice, it can prove costly to both marketers (whose algorithmic product offerings go unused) and to themselves (who fail to reap the benefits that algorithmic predictions often provide). In a departure from previous research focusing on *when* algorithm aversion proves more or less likely, we sought to identify and remedy one reason *why* it occurs in the first place. In seven pre-registered studies, we find that consumers tend to avoid algorithmic advice on the often faulty assumption that those algorithms, unlike their human counterparts, cannot learn from mistakes, in turn offering an inroad by which to reduce algorithm aversion: highlighting their ability to learn. Process evidence, through both mediation and moderation, examines why consumers fail to trust algorithms that err across a variety of prediction domains and how different theory-driven interventions can solve the practical problem of enhancing trust and consequential choice in algorithms.

KEYWORDS
learning from mistakes, algorithm appreciation, algorithm aversion, intervention, mistakes

## INTRODUCTION

Prediction technologies are rapidly altering the consumer landscape. From online shopping and at-home entertainment to internet search and medical diagnosis, prediction represents the frontier of consumer behavior. The ability to take data that is available (e.g., historical purchase patterns) and create data that is not (e.g., future purchase preferences) through the use of computational algorithms, machine learning, and artificial intelligence, provides new levels of depth and accuracy with which firms can cater their offerings to customers (Agrawal et al., 2018). The reach and pervasiveness of these technologies are continuing to change how individuals behave, interact, and make purchasing decisions.

To understand this ubiquity, one need only consider a typical morning routine. After walking out the door of their house, which is automatically lowering its temperature to save on heating bills while its occupants are away, our exemplar sets about driving to a coffee shop using a navigation app that provides up-to-the-second optimized driving directions making predictions based on the current and historical traffic patterns. This is all done while listening to a curated podcast or music playlist built from previous listening patterns. Upon arriving at the coffee shop, the facial recognition software on their phone makes unlocking it easier than typing 1–2–3-4, leaving their hands free to text message a co-worker to ask if they want a coffee, each spelling mistake automatically corrected, often before it can even be noticed. A quick tap of a payment app or credit card, which is verified immediately through predictive technology, ends the purchasing process. Finally, they can sit down to enjoy their coffee while reading a notification screen customized with a selection of news articles and other media, each curated based on personal data and the data of similar others, predicting what will be most interesting (and most likely to be clicked). As this example shows, as more and more data are used in analyzing

Accepted by Lauren Block, Editor; Associate Editor, Ann McGill

and predicting decisions, this process will only become more powerful, predictive, and pervasive.

As the marketplace moves from more traditional forecasting paradigms in which humans—whether friends, experts, or marketers—provide recommendations to those that rely on non-human algorithmic sources, it becomes increasingly important to understand the differences in how consumers interact with these different prediction sources. Far from an algorithmic takeover, research has documented so-called "algorithm aversion" or the general preference for humans' recommendations or predictions (Dawes, 1979; Dietvorst et al., 2015; Meehl, 1954; Promberger & Baron, 2006; Yeomans et al., 2019). However, this general preference can be overcome, as consumers appear open to the use of and reliance on algorithms under specific conditions (Logg et al., 2019), suggesting that algorithm aversion is by no means set in stone. Given the robust ability of algorithms to provide predictions that exceed the accuracy of human prediction (Dawes, 1979; Meehl, 1954), the present investigation asks not *which* domains best lend themselves to algorithm appreciation but, instead, *how* interventions can be leveraged in order to enhance algorithm appreciation (Dietvorst et al., 2018). In order to do so, the present investigation leverages a characteristic inherent to prediction—occasional error—to consider whether these inevitable mistakes might be framed not, in keeping with past research, as evidence of algorithm failure but instead as offering the potential for learning. Integrating the burgeoning literature on the benefits of mistakes (Reich et al., 2018; Reich & Maglio, 2020; Reich & Tormala, 2013), we propose that consumers are reluctant to trust algorithms that err but that those same errors, when seen as opportunities from which the algorithm can learn, enhance trust in and reliance on algorithms.

# THEORETICAL BACKGROUND

## Algorithmic versus human recommendation

Algorithms—statistical models, decision rules, or other mathematical procedures for forecasting—have been shown to consistently outperform humans in a variety of prediction tasks including academic performance, parole violation, graduate student success, and clinical diagnosis (Dawes, 1979; Dawes et al., 1989; Grove et al., 2000; Meehl, 1954). Despite this, a majority of research has shown that people prefer human predictions to those put forward by algorithms (Dana & Thomas, 2006; Dietvorst et al., 2015; Eastwood et al., 2015; Hastie & Dawes, 2009). This research also shows that people tend to favor human input over algorithmic input in decision-processes (Dietvorst et al., 2015; Önkal et al., 2009) and that they prefer that others, particularly professionals, seek out other human advice rather than algorithmic advice (Shaffer et al., 2013). Even when told that algorithms provide higher quality and more accurate predictions, people often continue to favor the suboptimal recommendations of humans. Research points to many reasons for this hesitancy toward algorithms, including the inability of algorithms to specify targets and account for individual differences (Grove & Meehl, 1996; Longoni et al., 2019) and concerns that the use of algorithms is unethical or dehumanizing (Dawes, 1979; Grove & Meehl, 1996).

However, as evidence continues to mount in favor of the efficacy and superiority of algorithmic predictions (over human forecasters), and as these predictions spread into new and diverse domains, consumers have become more comfortable in choosing algorithmic recommenders and, ultimately, in accepting and utilizing these predictions. Research in "algorithm appreciation," wherein algorithms are preferred over humans as sources of prediction, has provided evidence of changing preferences in favor of algorithms in different specific domains and scenarios (Dietvorst & Bharti, 2020; Dijkstra, 1999; Dijkstra et al., 1998; Logg et al., 2019). For instance, consumers appear more receptive to advice and forecasts stemming from an algorithm when the domain under consideration is objective rather than subjective (Castelo et al., 2019) and utilitarian rather than hedonic (Longoni & Cian, 2022). These developments can be seen as favorable insofar as prediction machines regularly offer greater accuracy than humans and will likely dominate the future of forecasting (Agrawal et al., 2018). Whereas this past research has identified the specific domains in which consumers are naturally more amenable to algorithms, the present investigation takes a different perspective in offering a theory-driven intervention—with two specific manifestations—designed to enhance consumer reliance on algorithms: learning from mistakes.

## Learning from mistakes

Even while supporting the contention that consumers fail to use superior algorithms in favor of inferior human forecasting, the paper that coined the phrase "algorithm aversion" in fact found that algorithmic forecasts were preferred in control conditions in which they were not seen to err (Dietvorst et al., 2015). In other words, consumers seem open to algorithms but with one caveat: that they not make mistakes. This turns out to be a substantial caveat, as the uncertainty inherent in forecasting the future makes it an error-filled enterprise. It is only when algorithms inevitably fall short of perfect accuracy that consumers dismiss them (Dietvorst et al., 2015), deriving from an inference that a single mistake signals that the algorithm is irrevocably flawed or simply broken (Dawes, 1979; Highhouse, 2008). Stubborn though it may be, this inference is often wrong, as algorithms can be capable of improvement and learning over time. If consumers believe that a mistaken algorithm is a broken algorithm, can this perception be remedied in order to bolster trust in algorithms?

Based on the heretofore separate literature on human mistake-making and learning from mistakes, the present investigation proposes that the answer to this question is yes. Just because they are largely unavoidable does not make mistakes desirable; from watching a bad movie to buying a lemon of a car to undergoing the wrong medical procedure, following mistaken forecasts and taking mistaken advice undermines consumer welfare. As a result, people tend to keep their mistakes to themselves (Edmondson, 1996; Stefaniak & Robertson, 2010; Uribe et al., 2002) for fear of being seen as incompetent and deserving of dismissal (Chesney & Su, 2010; Kunda, 1999; Palmer et al., 2010; Parker & Lawton, 2003) in much the same way that consumers dismiss mistaken algorithms. Still, the focal comparison in that work lies in consumer response to actors making or not making a mistake. In prediction, where errors and mistakes occur regularly and can thus be taken as a given, the more important issue might lie in how to manage or mitigate the overblown inferences of incompetence that often follow from mistakes.

When *people* make mistakes, observers often care more about how they respond to the mistake than about its occurrence in the first place. If people make an initial mistake and then keep making the same mistake repeatedly, observers rightly doubt the potential for change and growth. However, when people demonstrate that they have learned from their mistakes, making the initial mistake can prove not only not detrimental but, in fact, beneficial (Reich & Maglio, 2020). People who change their minds can be more persuasive than those who hold fast when observers see the change as resulting from new information and learning (Reich & Tormala, 2013). People who falter in pursuing a goal are seen as more likely to attain that goal than others who never falter, provided that they have corrected the original mistake (Kupor et al., 2018). Consumers who admit to past purchase mistakes in writing online reviews are more likely to be trusted than others who never made a mistake when the audience interprets response to the initial mistake as evidence of learning and gained expertise (Reich & Maglio, 2020). In each of these instances, acknowledging mistakes proves beneficial because observers appraise the mistake maker as having learned from the experience, interpreting her/his behavior through a lens of change and growth rather than seeing the mistake as diagnostic of a permanent, unfixable flaw (Dweck, 2011; Hong et al., 1995). Taken together, the findings from the literature on how people think about other humans who make mistakes suggest that, when framed as opportunities from which learning has occurred, mistakes foster trust.

## The present investigation

When observers see other people learn from mistakes, those observers have more confidence in mistake-makers and end up more willing to follow their advice. This occurs in large part because learning from mistakes serves as a reliable signal of gained expertise (Reich & Maglio, 2020), an element of credibility that causes people to be more persuasive in their communicative messaging (Berlo et al., 1969; Hovland et al., 1953; McGuire, 1978; Pornpitakpan, 2004). The construct of expertise allows us to build a conceptual bridge from humans who make mistakes to all entities that make mistakes. Research in human versus algorithmic prediction has shown that heightened perceptions of expertise can increase preferences for both algorithmic (Dijkstra, 1999; Dijkstra et al., 1998) and human (Arkes et al., 1986; Promberger & Baron, 2006) recommendations, as people prefer accurate forecasts and expertise is seen as a means by which to achieve it (Binzel & Fehr, 2013; Bolger & Wright, 1992; Hovland & Weiss, 1951; Sternthal et al., 1978). Similarly, Dietvorst et al. (2015) demonstrated that a loss of confidence in an algorithm's ability was the key factor that led to a preference for human predictions after an algorithm erred. However, in an early treatment of algorithm learning, Berger et al. (2021) conducted an experiment in which participants imagined working at a call center and had to forecast the number of incoming calls. For this objective, non-personal task, participants were more likely to rely on an algorithm when they watched (i.e., experienced) it improve over successive trials. Of greatest relevance to the present investigation, this finding provides preliminary evidence that algorithms are capable of bouncing back from prior missteps in the eyes of human observers questioning whether to place confidence and trust in them. Building from this prior work, the present investigation considers both objective and subjective domains of algorithmic prediction, non-personal and personal tasks, and manipulates algorithmic learning by having it described to participants rather than having them experience it.

Expertise predicts trust, including in the choice between reliance on human or algorithmic prediction, and research to date has suggested that both humans and algorithms are capable of being seen as having more of it than the other. However, research to date has not considered how one factor inherent to the forecasting process—the making of mistakes—might be leveraged as a means by which to enhance algorithm appreciation rather than result in unilateral algorithm aversion. This appears to stem from the fact that people see other people as capable of growth following mistakes, so, it stands to reason that they would place more stock in human predictions while eschewing predictions from algorithms that are seen as incapable of learning and growth (Dawes, 1979; Highhouse, 2008). However, as the literature on mistake-making suggests, mistakes garner trust when they act as springboards from which learning can occur. If human mistake-makers can vary in the extent to which observers see them as able to grow and learn or remain trapped in ineptitude, then perhaps algorithmic mistake-makers might also vary in the same manner, and with the same potential benefits.

The present investigation proposes that, just as not all human mistakes are seen as created equal, not all algorithmic mistakes should be seen as equally damning. Instead, algorithms should garner the least trust when they are seen as incapable of learning from their mistakes. This tends to be the standard assumption when consumers witness algorithms err, suggesting that different interventions and framings of algorithmic misprediction might overcome the default inference made by consumers. Study 1 first verifies this assumption, testing whether people perceive algorithms as less capable of learning than humans and whether these perceptions affect trust. To provide evidence of a robust effect across different prediction domains established by prior literature (Castelo et al., 2019), Study 1 documents this tendency for both objective and subjective predictions. Thereafter, Study 2 moves beyond learning in general to learning from mistakes more specifically by providing prior performance statistics (including successes and failures) for both humans and algorithms in a subjective task. By assessing both perceptions of the ability to learn from mistakes as well as trust, Study 2 tests the mediating role of perceived learning from mistakes on which prediction source—human or algorithm—people trust.

The remaining three studies examine the efficacy of different interventions designed to frame algorithms as capable of learning from mistakes. Study 3 manipulates not only prediction source type but also the inclusion (or absence) of performance statistics (similar to Study 2) designed to include learning evidence by making prior performance dynamic (i.e., improving over time). By introducing a key moderator (learning evidence), Study 3 provides support for our proposed process using a moderated mediation analysis to predict which prediction source people trust. Moving from trust to actual choice in an incentive-compatible design, Study 4 again uses the same prior performance intervention as Study 3 to provide support for the role of learning evidence in influencing consequential choice of an algorithm over a human. Study 5A introduces a second learning evidence intervention (what the algorithm is called), simultaneously comparing its trust-related effectiveness to our other learning intervention (learning performance), a traditional algorithm, and human prediction. Study 5B isolates its focus onto this more subtle manipulation of language to provide support for the role of learning in a different domain and with a different means by which to create a consequential choice setting. Finally, Study 6 provides a number of robustness checks for the level of performance and rate of improvement in the learning conditions as well as the importance of the decision with a different structure of incentive compatibility. Collectively, the studies test why consumers fail to trust algorithms that err across a variety of prediction domains as well as how different interventions can enhance trust and consequential choice. All of the studies report all manipulations and measures used, and each study was pre-registered.

# STUDY 1

As a foundation from which to build our investigation, Study 1 first replicates and extends a fundamental claim: that people perceive algorithms as less capable of learning compared to humans, by which they are as good as they can be out of the box and mistakes flag that they are fundamentally broken (Dawes, 1979; Highhouse, 2008). Providing a more modern update, Study 1 considers this question in the consumer domains examined by Castelo et al. (2019) and, accordingly, Study 1 includes not only a measure of learning ability (Dietvorst et al., 2015) but also the same trust measure as employed by Castelo et al. (2019). Because the latter investigation used two different domains (in the interest of examining its focal research question comparing algorithm aversion for objective and subjective forecasts), Study 1 also includes those same two domains (one objective and one subjective) in the interest of completeness. While we expect to replicate the basic effect (heightened algorithm trust in objective over subjective domains; Castelo et al., 2019), we more importantly hypothesize that participants will appraise algorithms as less capable of learning than humans across the different domains. This study was preregistered on AsPredicted.org (https://aspredicted.org/blind.php?x=zm2qm8).

## Method

We recruited a sample of two-hundred participants ($M_{age}$ = 38.47, $SD$ = 10.50; 51.5% female) from Amazon Mechanical Turk using CloudResearch. We utilized the CloudResearch Approved Participants feature, to ensure the participation of only high-quality participants who have passed CloudResearch's attention and engagement measures. Participants participated in exchange for monetary payment. Participants were randomly assigned to one of two domain conditions: objective or subjective. The objective domain consisted of recommending a disease treatment and the subjective domain consisted of predicting personality traits (see Castelo et al., 2019). We employed the same trust measure as Castelo et al. (2019). Specifically, in the subjective condition, we asked participants to indicate who they would trust more to predict personality traits. Participants responded on a 0 to 100 sliding bar (0 = *human psychologicst*, 50 = *no preference*, 100 = *algorithm*). In the objective condition, we asked participants who they would trust more to recommend disease treatment (0 = *human doctor*, 50 = *no preference*, 100 = *algorithm*). Participants were then asked, in random order, the perceived learning items adapted from Dietvorst et al. (2015) that measure human and algorithm learning. The two-item human perceived learning index ($r$ = 0.75, $p<0.001$) comprised of how much they thought humans can learn while performing a task like this and how much they thought humans could improve while performing a task like this. The two-item algorithm perceived learning index ($r$ = 0.88, $p<0.001$) was comprised of how much they thought algorithms can learn while performing a task like this and how much they thought algorithms could improve while performing a task like this. Participants responded on a scale ranging from 1 (*not at all*) to 7 (*very much*). Finally, participants completed demographic measures (gender and age).

## Results and discussion

### Trust

As predicted, trust in a human was higher than trust in an algorithm for both the subjective domain ($M$ = 25.50, $SD$ = 26.98), $t(97) = -8.99$, $p<0.001$, 95% CI = [−29.91, −19.09] and the objective domain ($M$ = 34.07, $SD$ = 28.64), $t(101) = -5.62$, $p<0.001$, 95% CI = [−21.56, −10.31]. Consistent with the findings of Castelo et al. (2019), we found that trust in the algorithm was higher in the objective domain ($M$ = 34.07, $SD$ = 28.64) than in the subjective domain ($M$ = 25.50, $SD$ = 26.98), $t(198) = 2.18$, $p$ = 0.031, $d$ = 0.31, 95% CI = [0.80, 16.33].

### Perceived learning index

As predicted, a paired sample t-test revealed that participants thought that a human was more capable of learning while performing the subjective task ($M$ = 5.54, $SD$ = 1.02) compared with an algorithm ($M$ = 4.69, $SD$ = 1.63), $t(100) = 4.18$, $p<0.001$, 95% CI = [0.45, 1.25]. The same pattern was observed for the objective task: a paired sample t-test revealed that participants thought that a human was more capable of learning while performing the objective task ($M$ = 5.78, $SD$ = 1.04) compared with an algorithm ($M$ = 4.60, $SD$ = 1.69), $t(97) = 6.26$, $p<0.001$, 95% CI = [0.81, 1.55].

Using modern consumer prediction domains, Study 1 thus updates and extends prior work. Conceptually replicating algorithm aversion (Dietvorst et al., 2015), participants generally placed less trust in algorithms than in humans. Conceptually replicating task-dependent algorithm aversion (Castelo et al., 2019), participants placed less trust in algorithms specifically charged with making a prediction for a subjective (versus an objective) domain. However, foundational to the present investigation, participants believed that humans can learn more than can algorithms, an effect that generalized across the different prediction domains.

# STUDY 2

Study 1 echoes and updates prior work suggesting that consumers believe that algorithms are less likely to learn in general. In the interest of providing more targeted evidence for our focal construct, Study 2 considers whether people think that algorithms are incapable in the more specific arena of learning from mistakes. Should participants see algorithms as less capable in this regard as well, it would strengthen the case for our particular goal in the subsequent studies: designing interventions to foster the perception that algorithms can learn from mistakes.

This is an issue of not only conceptual relevance but also applied importance in consideration of where many algorithms are gaining traction in the current marketplace: subjective domains (e.g., film, music, job search, dating apps, and psychological profiles). In subjective domains, we conjecture that mistakes—and learning therefrom—are particularly prevalent, manifested in everything from bad movie recommendations to bad partner pairings. While algorithms seem well positioned to rise to prominence in objective domains (meaning that consumer perceptions of trust should rise in kind), errors of the sort common in subjective domains need not be tantamount to dismissal of algorithms outright. Past research (Castelo et al., 2019) and Study 1 documented that mistrust of algorithms is higher for subjective than objective tasks, suggesting that this is the realm in which means by which to enhance trust of algorithms might be most important and with the most potential to observe improvement. Accordingly, Study 2 (and all subsequent studies) focus on subjective tasks.

Study 1 provided no information about the forecaster other than its type (human versus algorithm). In order to gain traction on the issue of learning from mistakes, Study 2 provides not only the type of forecaster but also performance statistics, which include a history of errors, for both humans and algorithms. We include this not only as a stepping stone from which we will build our intervention in subsequent studies but also because of the prevalence of this practice in the modern marketplace (e.g., online vendors touting the success rate of their past recommendations to customers and Netflix appending Percent Match ratings to titles as evidence-based estimates of successful pairings). By assessing perceptions of learning from those mistakes as well as trust in the forecaster, Study 2 will test the hypothesized mediating role of perceived learning from mistakes on trust of humans and of algorithms. This study was pre-registered on AsPredicted.org (https://aspredicted.org/blind.php?x=r9hz9y).

## Method

We recruited a sample of two-hundred participants ($M_{age}$ = 41.36, *SD* = 12.13; 51.3% female) from Amazon Mechanical Turk using CloudResearch. We utilized the CloudResearch Approved Participants feature, to ensure the participation of only high-quality participants who have passed CloudResearch's attention and engagement measures. Participants participated in exchange for monetary payment. Participants were randomly assigned to one of two agent conditions: human or algorithm. In the human condition, participants read:

> Below are the performance statistics of a psychologist who works for an online psychological service:
>
> 80% of the personality trait evaluations turned out to be correct, 20% were incorrect.

In the algorithm condition, participants read:

> Below are the performance statistics of an algorithm used in an online psychological service:
>
> 80% of the personality trait evaluations turned out to be correct, 20% were incorrect.

Next, participants were asked to indicate the extent to which they trust the [psychologist] algorithm to evaluate personality traits. Participants responded on a scale ranging from 1 (*not at all*) to 7 (*very much*). They then completed a two-item perceived learning from mistakes measure ($r$ = 0.86, $p$<0.001) adapted from Tjosvold et al. (2004): "How much do you think the [psychologist] algorithm can learn from mistakes?" and "How much do you think the [psychologist] algorithm can improve following a mistake?" Again, participants responded on a scale ranging from 1 (*not at all*) to 7 (*very much*). Finally, participants completed demographic measures (gender and age).

## Results and discussion

### Trust

As predicted, trust in the human to evaluate personality traits was higher (*M* = 4.99, *SD* = 0.96) than trust in the algorithm (*M* = 4.62, *SD* = 1.16), $t$(191.13) = 2.46, $p$ = 0.015, $d$ = 0.35, 95% CI = [0.07, 0.67].

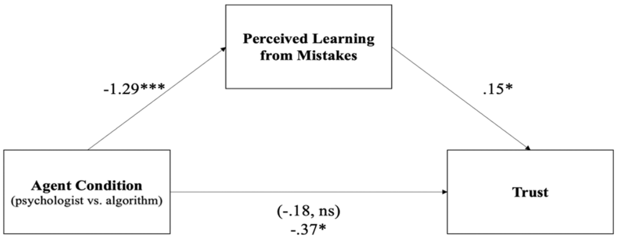
Perceived Learning from Mistakes
-1.29***
.15*
Agent Condition (psychologist vs. algorithm)
Trust
(-.18, ns)
-.37*

FIGURE 1 Study 2 mediation analysis. *Notes*. The path coefficients are unstandardized betas. The values in parentheses indicate the effect of condition on the dependent variable after controlling for the mediator. ***$p$<0.001, **$p$<0.01, *$p$<0.05.

### Perceived learning from mistakes index

As predicted, participants perceived the human as more capable of learning from their mistakes (*M* = 5.77, *SD* = 1.02) than the algorithm (*M* = 4.48, *SD* = 1.56), $t$(170.50) = 6.92, $p$<0.001, $d$ = 0.98, 95% CI = [0.92, 1.65].

*Mediation analysis*. To test whether differences in perceived learning from mistakes mediate the effect of agent condition on trust, we conducted a mediation analysis with 5000 bootstraps following the procedures recommended by Hayes (2013), in which condition was entered as the independent variable, perceived learning from mistakes index was entered as the mediator, trust was entered as the dependent variable. Consistent with our theory, compared with the algorithm, the human was perceived as more capable of learning from their mistakes which in turn resulted in increased trust, 95% CI for the indirect effect: [−0.3862, −0.0170]; see Figure 1.

These results indicate that the learning limitations consumers expect of algorithms are not restricted to general terms (Study 1). Instead, participants in Study 2 learned that a human or an algorithm had erred at a given rate, but those mistakes were appraised differently as a function of who made the errors. When the forecast came from an algorithm, participants saw the errors as something from which little learning could occur, accounting for the dampened trust in the algorithm relative to a human forecaster. After documenting a prevalent and meaningful phenomenon in the first two studies, our remaining studies consider how different interventions might mitigate its occurrence and downstream consequences.

# STUDY 3

Study 2 used prior performance to expose participants to information about the performance history of the predicting agent (human or algorithm). Those static snapshots offered no evidence of learning in and of themselves; instead, participants brought their own expectations to bear in interpreting this data, apparently presuming that the human could learn better than could the algorithm. However, should those performance histories be dynamic rather than static, they would be able to offer evidence of learning should the performance of the agent improve over time. That is, rather than a summary snapshot of past performance, the history might be presented in a manner that reflects an increasingly positive trend (Kupor & Laurin, 2020; Maglio & Polman, 2014, 2016; Reich et al., 2021b), allowing participants to witness, directly, a retrospective sense of learning by the agent over successive mistakes (growing smaller in number and rate).

Study 3 integrates this logic as one particular application of our broader construct of interest (learning from mistakes) that allows for the design of an intervention to augment perceived algorithm learning. Thus, the design of Study 3 builds from the performance history methodology used in Study 2, but with the inclusion of a dynamic component by which the predictions of the agent reflect learning. Accordingly, this study manipulates not only agent type (human or algorithm) but also whether the performance history statistics include learning evidence or not (i.e., a pattern representative of an improving trajectory). Our hypothesis development (with supporting evidence from Study 2) has suggested that humans are seen as innately capable of learning from mistakes, suggesting that only algorithms should receive a boost in trust as a result of making prior learning salient. Accordingly, the design of Study 3 allows us to examine the effect of providing learning evidence (versus no such evidence) on trust via perceptions of an ability to learn as a function of prediction agent. We predict that, in the absence of evidence of learning, algorithms will be seen as less capable of learning than humans, accounting for lower trust. However, when provided with evidence of learning, we predict that algorithms will be seen as equally capable of learning as humans, resulting in a level of trust commensurate with that of humans as a result of this learning-based intervention. We test these hypothesized relationships using a moderated mediation analysis. This study was pre-registered on AsPredicted.org (https://aspredicted.org/blind.php?x=76cu5n).

## Method

We recruited a sample of four-hundred participants ($M_{age}$ = 40.33, *SD* = 12.42; 51.4% female) from Amazon Mechanical Turk using CloudResearch. We utilized the CloudResearch Approved Participants feature, to ensure the participation of only high-quality participants who have passed CloudResearch's attention and engagement measures. Participants participated in exchange for monetary payment. Participants were randomly assigned to one condition in a 2 (agent: human vs. algorithm) x 2 (performance data: with learning evidence vs. without) between-subjects design. Largely resembling the setup of Study 2, participants in the without learning evidence conditions read:

> Below are the performance statistics of an algorithm used in [a psychologist who works for] an online psychological service (% of personality trait evaluations that turned out to be correct or incorrect):
>
> 80% correct, 20% incorrect.

In the learning evidence conditions, participants read:

> Below are the first year performance statistics of an algorithm used in [psychologist who works for] an online psychological service (% of personality trait evaluations that turned out to be correct or incorrect):
>
> First 3 months: 60% correct, 40% incorrect.
>
> First 6 months: 70% correct, 30% incorrect.
>
> First year: 80% correct, 20% incorrect.

Next, participants completed the same one-item trust measure and the same two-item perceived learning from mistakes measure ($r = 0.86$, $p < 0.001$) as in Study 2. Finally, participants completed demographic measures (gender and age).

## Results and discussion

### Trust

We submitted the trust data to a 2 (agent: human vs. algorithm) x 2 (learning evidence: with vs. without) analysis of variance (ANOVA). This analysis revealed a main effect of agent, $F(1, 396) = 4.04$, $p = 0.045$, $\eta_p^2 = 0.01$ and a main effect of learning evidence, $F(1, 396) = 5.62$, $p = 0.018$, $\eta_p^2 = 0.01$. Importantly, these main effects were qualified by a significant interaction between agent and performance data, $F(1, 396) = 5.19$, $p = 0.023$, $\eta_p^2 = 0.01$. As illustrated in Figure 2, participants trusted the algorithm significantly more when performance data included learning evidence ($M = 5.00$, $SD = 1.03$) compared with when performance data did not include learning evidence ($M = 4.48$, $SD = 1.34$), $F(1, 396) = 10.85$, $p = 0.001$, $\eta_p^2 = 0.03$. Conversely, there was no difference in trust of a human when performance data included learning evidence ($M = 4.97$, $SD = 1.03$) and when it did not ($M = 4.96$, $SD = 1.04$), $F(1, 396) = 0.004$, $p = 0.948$.

### Perceived learning from mistakes index

We submitted the perceived learning from mistakes index data to a similar 2 (agent: human vs. algorithm) x 2 (learning evidence: with vs. without) analysis of variance

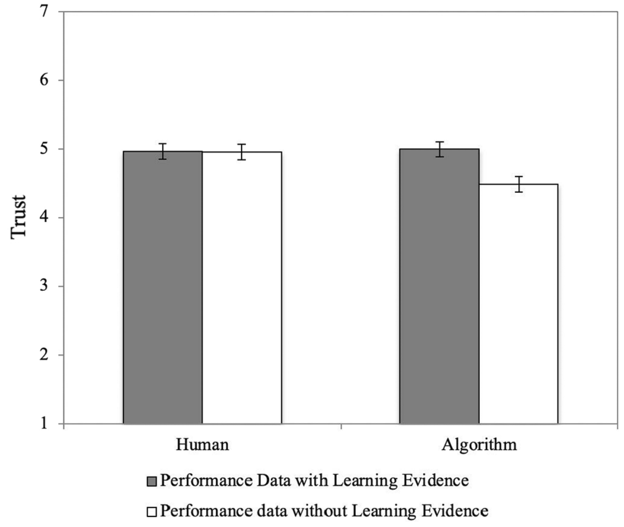

**FIGURE 2** Trust as a function of agent and learning evidence, Study 3

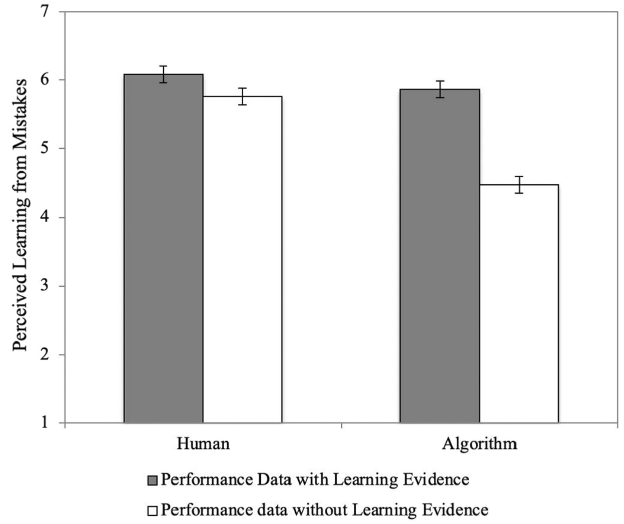

**FIGURE 3** Perceived learning from mistakes as a function of agent and learning evidence, Study 3

(ANOVA). This analysis revealed a main effect of agent, $F(1, 396) = 37.68$, $p<0.001$, $\eta_p^2 = 0.09$ and a main effect of learning evidence, $F(1, 396) = 49.43$, $p<0.001$, $\eta_p^2 = 0.11$. Importantly, these main effects were qualified by a significant interaction between agent and learning evidence, $F(1, 396) = 19.23$, $p<0.001$, $\eta_p^2 = 0.05$. As illustrated in Figure 3, participants thought the algorithm was significantly more capable of learning from mistakes when performance data included learning evidence ($M = 5.87$, $SD = 0.90$) compared with when performance data did not include learning evidence ($M = 4.48$, $SD = 1.63$), $F(1, 396) = 65.49$, $p<0.001$, $\eta_p^2 = 0.14$. Conversely, there was only a marginal difference in perceptions of learning from mistakes when performance data of the human included learning evidence ($M = 6.08$, $SD = 1.04$) compared with when it did not ($M = 5.76$, $SD = 1.19$), $F(1, 396) = 3.48$, $p = 0.063$, $\eta_p^2 = 0.01$.

## Moderated mediation analysis

To test our proposed process, we ran a moderated mediation with 5000 bootstraps (Model 7 in Process, Hayes, 2013), with agent as the independent variable, the perceived learning from mistakes index as the mediator, learning evidence as the moderator, and trust as the dependent variable. As predicted, the model revealed a significant moderated mediation, 95% CI for the index of moderated mediation: [0.2481, 0.7109]. When performance data did not include learning evidence, the human was perceived as being more capable of learning from mistakes than the algorithm, which in turn drove trust, 95% CI for the indirect effect: [−0.7759, −0.3600]. In contrast, when performance data included learning evidence, the mediation model revealed no difference between the human and the algorithm, 95% CI for the indirect effect: [−0.2106, 0.0280].

Study 3 provides novel, important evidence that perceptions of algorithms as unable to learn are malleable and not fixed. For the conditions in which no learning evidence was provided, participants appeared to rely on their assumptions about the innate ability of the agents: Humans can learn but algorithms cannot, driving a divergence in trust that conceptually replicates past research. However, for the conditions in which learning evidence was provided (i.e., in which we manipulate our proposed mediator), all downstream differences between humans and algorithms were eliminated vis-a-vis boosts in perceived learning and trust for the algorithm. By providing past performance statistics that comprise a history of improvement, participants appear capable of overcoming their concern that algorithms cannot learn, as manifested by the effect on our conceptual mediator, perceptions of learning ability. From there, consistent with the tendency for consumers to trust and follow the advice of agents who demonstrate expertise by learning from their past mistakes (Reich & Maglio, 2020), participants proved more willing to trust algorithms that learn. Indeed, appearing capable of learning from mistakes (at least in the form of performance history) made participants in Study 3 just as trusting of algorithms as of human forecasters (who did not appear to need a learning history to be trusted).

# STUDY 4

Study 3 implemented a small but important change in how performance statistics were described: Beyond the static snapshots of Study 2, presenting an improving pattern of performance proved sufficient to make consumers believe that algorithms were capable of learning. As a result, participants in Study 3 trusted algorithms to a degree commensurate with human forecasters, suggesting that learning evidence can change judgment. But can it also change choice? Study 4 sought to test this question.

We first constructed three algorithm types: a control algorithm (providing no information about the algorithm other than that it was an algorithm), an algorithm without learning evidence, or an algorithm with learning evidence. Each algorithm type was paired with the same human agent to create three experimental conditions, and participants were tasked with choosing whether to accept the advice provided by the human or the algorithm described as a function of their experimental condition. Thus, Study 4 asks participants to make a forced choice between a human forecaster or an algorithm; we predict that choice share in favor of the algorithm will be greatest among participants for whom the algorithm has learning evidence. Switching from a judgment (i.e., of trust, as in Studies 1–3) to choice also allowed Study 4 to adopt an incentive compatible design: The choice made by participants to opt for the human or the algorithm carried consequential relevance, as making the better choice came with a financial incentive. Thus, Study 4 provides evidence for the role of learning evidence on influencing actual choice. This study was pre-registered on AsPredicted.org (https://aspredicted.org/blind.php?x=jm77br).

## Method

We recruited a sample of three-hundred participants ($M_{age}$ = 39.64, $SD$ = 11.94; 54.7% female) from Amazon Mechanical Turk using CloudResearch. We utilized the CloudResearch Approved Participants feature, to ensure the participation of only high-quality participants who have passed CloudResearch's attention and engagement measures. Participants participated in exchange for monetary payment. In all conditions, participants were told that they will be asked to choose between a psychologist's personality evaluation and an algorithm's personality evaluation of another participant. To make choosing correctly feel consequential, all participants were told that at the end of the study, we would score the accuracy of the psychologist and the algorithm against a standardized personality measure to code for accuracy and those who chose the accurate personality evaluation—psychologist or algorithm—would be entered into a lottery to win an additional $1 bonus. In fact, all participants received the $1 bonus at the conclusion of the experimental session.

Participants were randomly assigned to one of three conditions with varying descriptions of the algorithm: control, algorithm performance data without learning evidence, or algorithm performance data with learning evidence. In the control condition, no information was provided about the algorithm other than that it was an algorithm. The two performance data algorithms were described in a manner identical to Studies 2 and 3: The algorithm without learning evidence was described as having a performance history of 80% success and 20% failure; the algorithm with learning evidence was described similarly but with that performance having improved over the past year. In all conditions, for the human agent option, no information was provided about the human other than that it was a human. They then indicated their choice by selecting either "The psychologist" radio button or "The algorithm" radio button. Finally, participants completed demographic measures (gender and age).

## Results and discussion

### Choice

Two dummy variables were created: one for the algorithm without learning evidence condition (0 = control, 1 = without learning evidence, 0 = with learning evidence) and one for the control condition (1 = control, 0 = without learning evidence, 0 = with learning evidence) to allow for comparison with the algorithm with learning evidence condition. The dependent variable of choice was coded as 1 = human and 2 = algorithm. As predicted, a binary logistic regression revealed that participants in the algorithm with learning evidence condition were more likely to choose the algorithm (66.3%) than participants in the algorithm without learning evidence condition (50.5%; $b = -0.66$, SE = 0.29, $p = 0.024$, 95% CI = [0.292, 0.918]) and participants in the control condition (26.7%; $b = -1.69$, SE = 0.31, $p<0.001$, 95% CI = [0.101, 0.340]). In addition, participants in the algorithm without learning evidence condition were more likely to choose the algorithm (50.5%) than participants in the control condition (26.7%; $b = 1.03$, SE = 0.30, $p = 0.001$, 95% CI = [1.552, 5.036]).

In the algorithm with learning evidence condition, the choice of the algorithm (66.3%) was significantly bigger than the choice of the human (33.7%), $\chi2(1) = 10.45$, $p = 0.001$, $d = 0.69$. In the without learning evidence condition, the choice of the algorithm (50.5%) was not significantly different from the choice of the human (49.5%), $\chi2(1) = 0.10$, $p = 0.921$, $d = 0.02$. Finally, in the control condition, the choice of the algorithm (26.7%) was significantly smaller than the choice of the human (73.3%), $\chi2(1) = 21.87$, $p<0.001$, $d = 1.05$.

These results offer new insights about how consumers decide whether to choose a human or an algorithm to provide the most accurate forecast. The control condition conceptually replicates algorithm aversion (Dietvorst et al., 2015), with participants actively avoiding the algorithm. However, when merely provided with a performance history designed to include no evidence of learning, the rate at which participants chose the human or the algorithm was roughly at chance—evidence for what might be termed "algorithm indifference." We note that, in Studies 2 and 3, the no-learning conditions were contrasted only against the learning conditions. Study 4, however, includes a control condition, allowing for a comparison between performance history with no learning and a pure control condition (no information about performance history or learning provided). Here, it appears that any mention of the performance of an algorithm garners more consumer trust in that algorithm than when that information is not provided. We return to the implications of this unanticipated result in the General Discussion. To our primary prediction, we close in highlighting that Study 4 provided support for the effectiveness of learning evidence on consumer trust. Even in an incentive-compatible context, participants overcame any potential algorithm aversion and behaved in a manner more consistent with algorithm appreciation—indeed, algorithm investment—by choosing it at a higher rate.

## STUDY 5A

Having provided evidence for an effect of perceived algorithm learning on judgments of trust (Study 3) and choice (Study 4), the current study seeks to speak to the breadth of this effect in two notable ways. First, Study 5A provides evidence for the benefit of learning evidence in a novel domain (online dating), testing a noteworthy robustness check. Second, the means by which we manipulated learning in Studies 3 and 4 presented more pieces of performance information (success rates at 3 different time points) than in the non-learning conditions, raising the possibility that this sheer amount of information made different characteristics more salient to participants. Study 5A examines the effectiveness of a novel learning intervention based on language rather than performance history.

The literature on psycholinguistics documents the many ways in which the names given to products change how consumers think about and behave toward those products (e.g., Maglio et al., 2014; Maglio & Feder, 2017; Rabaglia et al., 2016; Shrum & Lowrey, 2007; Yorkston & Menon, 2004). Accordingly, Study 5A considers whether merely changing what the algorithm is called might impact the extent to which consumers trust it. Toward this goal, the current consumer landscape offered a clear candidate: machine learning. While there are definite points of difference between algorithms and machine learning, all machine learning is underpinned by algorithms; though not all algorithms utilize machine learning, any algorithm that refines and improves its forecasting capability through trial and error can be conceptualized as machine learning (Mitchell, 1997). Because our investigation targets learning from mistakes as applied to algorithms, we believe it is a fair characterization to describe the algorithms under consideration with the current set of studies as machine learning. Accordingly, it becomes a matter of arbitrary semantic choice as to whether to describe the agent as an algorithm or as a machine learning algorithm. However, we propose that the consequences of this choice will prove far from arbitrary, as the learning signaled by the latter terminology should enhance consumer trust in a manner consistent with our hypotheses. To allow for a full comparison of our different interventions, Study 5A includes both the novel "machine learning" naming intervention as well as the learning evidence intervention from our earlier studies as well as a traditional (control) algorithm and a human forecaster. This study was pre-registered on AsPredicted.org (https://aspredicted.org/blind.php?x=a3su25).

### Method

We recruited a sample of four-hundred participants ($M_{age}$ = 38.50, *SD* = 11.64; 48.5% female) from Amazon Mechanical Turk using CloudResearch. We utilized the CloudResearch Approved Participants feature, to ensure the participation of only high-quality participants who have passed CloudResearch's attention and engagement measures. Participants participated in exchange for monetary payment. Participants were randomly assigned to one of four agent conditions: human, traditional analytical algorithm, machine learning algorithm, or learning evidence algorithm. In the human condition, participants read:

> An online dating company can send you romantic partner recommendations. Below are last year's performance statistics of a psychologist who works for the online dating company: 85% good match, 15% bad match.

In the traditional analytical algorithm condition, participants read:

> An online dating company can send you romantic partner recommendations. Below are last year's performance statistics of a traditional analytical algorithm used in the online dating company: 85% good match, 15% bad match.

In the machine learning algorithm condition, participants read:

> An online dating company can send you romantic partner recommendations. The company uses a machine learning algorithm, with each mistake, the algorithm learns more, allowing it to provide better recommendations. Below are last year's performance statistics of the machine learning algorithm used in the online dating company: 85% good match, 15% bad match.

In the learning evidence algorithm condition, participants read:

> An online dating company can send you romantic partner recommendations. Below are last year's performance statistics of the algorithm used in the online dating company:
>
> First 3 months: 65% good match, 35% bad match.
>
> First 6 months: 75% good match, 25% bad match.
>
> First year: 85% good match, 15% bad match.

Next, participants completed the same one-item trust measure as in the previous studies. Finally, participants completed demographic measures (gender and age).

## Results and discussion

### Trust

As predicted, a one-way ANOVA revealed a significant effect of agent condition on trust, $F(3, 396) = 4.80$, $p = 0.003$. Planned contrasts revealed that participants trusted the human agent significantly more ($M = 5.00$, $SD = 1.14$) than they trusted the traditional analytical algorithm ($M = 4.47$, $SD = 1.44$), Fisher's LSD: $p = 0.003$, $d = 0.41$, 95% CI = [0.18, 0.87]. There was no difference in trust between the human agent and the machine learning algorithm ($M = 5.05$, $SD = 1.11$), Fisher's LSD: $p = 0.774$, 95% CI = [−0.39, 0.29], or the learning evidence algorithm ($M = 4.99$, $SD = 1.21$), Fisher's LSD: $p = 0.954$, 95% CI = [−0.33, 0.35]. The latter two did not differ from each other, Fisher's LSD: $p = 0.731$, 95% CI = [−0.28, 0.40]. Trust in the traditional analytical algorithm was significantly lower than trust in the machine learning algorithm, Fisher's LSD: $p = 0.001$, 95% CI = [0.23, 0.92], as well as trust in the learning evidence algorithm Fisher's LSD: $p = 0.003$, 95% CI = [0.17, 0.86]. Figure 4 summarizes these results.

With this study, we simultaneously document the effectiveness of the two interventions put forth by the present investigation. Conceptually replicating Studies 3 and 4 in the novel domain of online matchmaking, the learning evidence algorithm saw a level of trust commensurate with the human agent while the traditional algorithm again received the lowest trust ratings, consistent with algorithm aversion. Study 5A added a new intervention that did not require explicit evidence of learning history in the form of statistics but merely the inclusion of the word learning in the name of the algorithm ("machine learning algorithm") and a definition thereof, yet consumer trust in this prediction agent was just as high as it was in the former two conditions. Next, Study 5B considers whether this new intervention might reliably impact consequential decisions in a manner akin to the more explicit learning evidence impacting a consequential decision in Study 4. Furthermore, Study 5B uses an even subtler language-based intervention in that Study 5A stated not only that the algorithm was a "machine learning algorithm" but also provided a complete definition of what that meant. Would the effect generalize by using just the language-based intervention without including the definition?

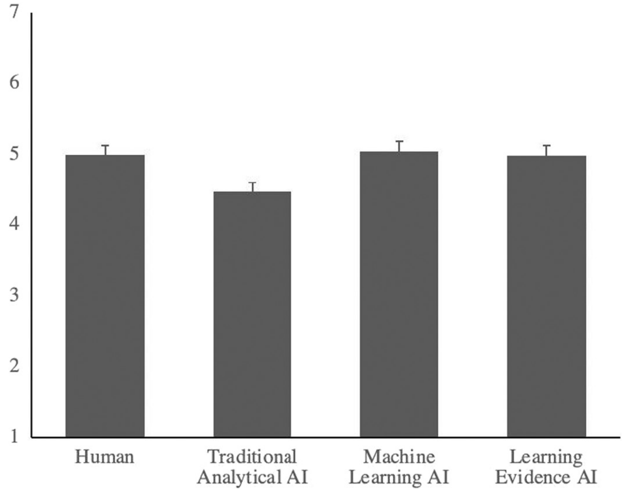

**FIGURE 4** Trust as a function of agent condition, Study 5A

# STUDY 5B

Study 5A found that changing the name applied to an algorithm shifted trust in it, and Study 5B aims to use this subtle manipulation of perceived algorithm learning in order to both enhance the ecological validity and the possible application of the current work. In it, participants are asked to make a consequential choice of whether to rely on their own judgment or on that of an algorithm. This methodological change from several of our previous experiments allows Study 5B to pit preference for algorithms against not some undefined other human but against human choosers themselves. Further, Study 5A used the terminology "traditional analytical algorithm" and found decreased trust in it. Though we chose it to match the word length of the condition describing a "machine learning algorithm," that phrasing may have inadvertently made the algorithm seem old or outdated, accounting for the reduced trust we observed. Accordingly, the experimental conditions in Study 5B use the more straightforward terms "algorithm" and "machine learning algorithm." In the interest of robustness testing, Study 5B applies these terms to examine our focal question in a new domain: art. By pivoting to art, Study 5B tests our predictions in a new incentive-compatible manner in order to complement the incentive-compatible design of Study 4. This study was pre-registered on AsPredicted.org (https://aspredicted.org/5HZ_WGH).

## Method

We recruited a sample of two-hundred participants ($M_{age}$ = 40.76, *SD* = 11.71; 49% female) from Amazon Mechanical Turk using CloudResearch. We utilized the CloudResearch Approved Participants feature, to ensure the participation of only high-quality participants who have passed CloudResearch's attention and engagement measures. Participants participated in exchange for monetary payment. Participants were randomly assigned to one of two agent conditions: algorithm or machine learning algorithm. Participants were told that the task involved estimating the quality of a piece of art. They could either use their own estimation or the estimation of either an "algorithm" or a "machine learning algorithm." To make choosing correctly feel consequential, all participants were told that, at the end of the study, we would score the accuracy of the estimation compared with the estimates of an expert art curator. With this benchmark for estimation accuracy, we would conduct a lottery among the participants who provided an accurate estimate, and the winner would receive a 10-cent bonus. In fact, all participants received the 10-cent bonus at the conclusion of the experimental session.

Specifically, participants were told "The estimation you provide today will be compared to the estimation of an art curator. We will consider an accurate estimation as one that is within 10 points of the art curator's estimation. The estimation will be on scale from 0 = low quality to 100 = high quality. Please look at the artwork." Participants were shown an art piece and were asked to indicate who they would like to estimate the quality of the art piece by selecting one of the following two radio buttons: "I would like to estimate;" "I would like the [machine learning] algorithm to estimate." They were then thanked for their selection and were notified that they would receive the 10-cent bonus as a token of our appreciation (i.e., all participants received the bonus regardless of their choice). Finally, participants completed demographic measures (gender and age).

## Results and discussion

### Choice

As predicted, a chi-square analysis revealed that when the algorithm was described as a "machine learning algorithm," participants were more likely to choose it over themselves to make the art quality estimation (66.7%), whereas when the algorithm was labeled solely as an "algorithm," participants were less likely to choose it over themselves to make the art quality estimation (43.6%; $\chi2(1)$ = 10.78, *p* = 0.001, Cramer's V = 0.232).

Study 5B demonstrated that subtle manipulation of perceived algorithm learning (labeling the algorithm as a "machine learning algorithm") can shift consumer preferences toward the algorithm in a consequential choice setting. That is, even in the absence of explicit information about past performance that demonstrates *how* an algorithm has learned from mistakes in the past, simply denoting *that* an algorithm is capable of learning from mistakes proves sufficient to overcome algorithm aversion.

# STUDY 6

Our final study, like Study 5B, stays in the domain of art but makes several important changes to test the robustness of our hypotheses. First, after having documented the effectiveness of a linguistic intervention in Studies 5A and 5B, Study 6 returns to the reporting of past performance statistics. We make this change because earlier studies reported rates indicative of high performance (80% success) and quick learning (an improvement of 10 percentage points over successive periods) that, together, might have been seen as approaching near perfection. Accordingly, Study 6 operationalizes both lower overall performance and slower learning (in the learning conditions) while holding performance constant. Second, some of our previous experiments were incentive compatible while others were not. With Study 6, we introduce incentive-compatibility as a between-subjects experimental factor in order to consider whether the stakes (i.e., importance or consequences of an error) for the decision might moderate our core findings. For example, would high (versus low) stakes make algorithm aversion more likely, causing consumers to eschew algorithms regardless of their ability to learn? Third, and toward this objective, Study 6 introduces yet another means by which to operationalize incentive compatibility. This study was pre-registered on AsPredicted.org (https://aspredicted.org/DNW_3YN).

## Method

We recruited a sample of eight-hundred participants ($M_{age}$ = 40.33, *SD* = 12.42; 51.4% female) from Amazon Mechanical Turk using CloudResearch. We utilized the CloudResearch Approved Participants feature, to ensure the participation of only high-quality participants who have passed CloudResearch's attention and engagement measures. Participants participated in exchange for monetary payment. Participants were randomly assigned to one condition in a 2 (agent: human vs. algorithm) x 2 (performance data: with learning evidence vs. without) x 2 (stakes: high vs. low) between-subjects design. In the high stakes conditions, participants read "Welcome to our art study. In today's study, you might be eligible to receive an art piece that will either be selected for you by an algorithm or by an art curator. 10% of all participants will actually receive the art piece selected in this study." In the low stakes conditions, participants read "Welcome to our art study! In today's study, you will be asked to evaluate an art piece that will either be selected for you by an algorithm or by an art curator." Thus, all participants chose whether to have an art curator or an algorithm choose a piece of art for them, but the stakes varied in that they would either simply evaluate that piece of art (low stakes) or possibly receive a physical copy of that piece of art (high stakes). On the next page, participants in the without learning evidence conditions read:

> Below are the two-year performance statistics of an algorithm used in [an art curator who works for] an online art gallery (% artworks that sold versus did not sell):
>
> 65% sold, 35% did not sell.

In the learning evidence conditions, participants read:

> Below are the two-year performance statistics of an algorithm used in [an art curator who works for] an online art gallery (% artworks that sold versus did not sell):
>
> 2 years ago: 55% sold, 45% did not sell.
>
> Last year: 60% sold, 40% did not sell.
>
> So far this year: 65% sold, 35% did not sell.

Next, participants were asked to indicate their preference using a measure adapted from Castelo et al. (2019): "Who would you prefer to select an art piece for you?" Participants reported their responses on a 0 to 100 sliding scale, with 0 labeled as art curator, 50 labeled as no preference, and 100 labeled as algorithm. Regardless of their response, all participants were then shown the same art piece. In order to maintain the cover story, all participants were asked to what extent the art piece was impressive on a scale from 1 (*not at all*) to 7 (*extremely impressive*), and those in the high-stakes condition were told that it was the piece of which they might win a physical copy. Finally, participants completed demographic measures (gender and age). At the conclusion of the study, 10% of the participants in the high stakes condition were shipped a poster of the art piece.

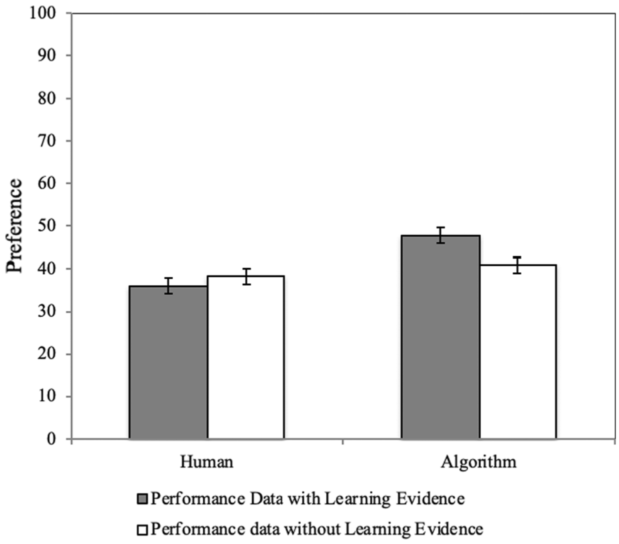

**FIGURE 5** Preference as a function of agent and learning evidence, Study 6

## Results and discussion

### Preference

We submitted the preference data to a 2 × 2 × 2 ANOVA with agent, performance data, and stakes as the independent variables. This analysis revealed a main effect of agent, $F(1, 794) = 14.89$, $p < 0.001$, $\eta_p^2 = 0.02$, and no main effect of performance data, $F(1, 794) = 1.58$, $p = 0.209$, or stakes, $F(1, 794) = 2.10$, $p = 0.147$. Importantly, there was a significant two-way interaction between agent and performance data, $F(1, 794) = 5.78$, $p = 0.016$, $\eta_p^2 = 0.01$. None of the other two-ways interactions ($p$s>0.55) or the three-way interaction were significant, $F(1, 794) = 0.53$, $p = 0.468$. Accordingly, we collapsed across the stakes factor in our analyses.

As illustrated in Figure 5, preference for the algorithm was significantly higher when evidence of learning data was provided ($M = 47.73$, $SD = 25.81$) compared with when performance data did not include learning evidence ($M = 40.95$, $SD = 26.32$), $F(1, 798) = 6.67$, $p = 0.010$, $\eta_p^2 = 0.01$. Conversely, there was no difference in preference for the human when performance data included learning evidence ($M = 36.07$, $SD = 26.99$) and when it did not ($M = 38.32$, $SD = 26.15$), $F(1, 798) = 0.731$, $p = 0.393$.

As part of the cover story, participants were asked to indicate how impressive they thought the art piece was. For completeness of data reporting, we submitted the data to a 2 x 2 x 2 ANOVA with agent, performance data, and stakes as the independent variables. None of the two-way interactions ($p$s>0.723) or the three-way interaction were significant ($p = 0.864$). This analysis revealed a main effect of agent, $F(1, 792) = 6.87$, $p = 0.009$, $\eta_p^2 = 0.01$, such that when the art piece was ostensibly selected by a human, it was perceived as significantly more impressive ($M = 4.18$, $SD = 1.85$) than when it was selected by an algorithm ($M = 3.85$, $SD = 1.81$). There was no main effect of performance data, $F(1, 792) = 1.51$, $p = 0.219$ or stakes, $F(1, 792) = 0.002$, $p = 0.961$.

Study 6 thus documents the robustness of our effect across several variations to our experimental design. Even with a lower overall performance and a slower rate of learning, algorithms capable of learning witness an increase in choice share relative to algorithms that do not. This effect of the mere capacity for learning is consistent with the results from Studies 5A and 5B. Additionally, the importance or consequences of the decision have a negligible impact on this overall effect, as Study 6 supported our hypothesis for both low- and high-stakes decisions between a human and an algorithm. Our high-stakes conditions provide evidence for our effect in a novel realm of incentive-compatibility. Rather than providing a financial incentive for correctly choosing between a human and an algorithm, the high stakes of Study 6 entailed participants making a product-related choice under circumstances in which they actually stood to receive that product.

# GENERAL DISCUSSION

A popular 2015 book asked in its title "What to Think About Machines That Think" (Brockman, 2015). As improvements in artificial intelligence, machine learning, and algorithmic forecasting continue to spread into new and diverse prediction domains, the need to understand what consumers think about these machines and their outputs—recommendations, predictions, and forecasts—continues to grow in kind. The inevitable rise of prediction machines, coupled with their tendency to outperform human forecasters, will necessitate not only a deeper understanding of the circumstances under which consumers opt for one agent over the other but also a means by which to leverage these insights to design interventions that help consumers help themselves in the form of placing well-earned trust in algorithmic forecasts.

Our research contributes to marketing and consumer behavior by speaking to both issues. Through the lens of mistake-making, seven experiments find that consumers tend to avoid algorithmic advice on the often faulty assumption that those algorithms, unlike their human counterparts, cannot learn from errors, in turn offering an inroad by which to remedy algorithm aversion. Study 1 replicated task-dependent algorithm aversion (Castelo et al., 2019), by showing that participants placed less trust in algorithms in a subjective (versus an objective) domain. Central to the current investigation, participants further believed that humans are more capable of learning than algorithms, an effect that generalized across both objective and subjective domains. Study 2 provided more targeted evidence by examining learning from mistakes. We found that perceived learning from mistakes mediated the effect of agent (human or algorithm) on trust. In Study 3, we introduced a key moderator (learning evidence), to provide evidence for our proposed process using a moderated mediation analysis. In the absence of evidence of learning, algorithms were seen as less capable of learning than humans, accounting for lower trust. However, when provided with evidence of learning, algorithms were seen as equally capable of learning as humans, resulting in a level of trust commensurate with that of humans. Study 4 moved from trust to actual choice in an incentive-compatible design, providing evidence for the role of our learning evidence intervention in influencing consequential choice of an algorithm over a human. Study 5A included a full comparison of our different interventions. We examined both a novel machine learning naming intervention as well as the learning evidence intervention from our earlier studies as well as a traditional (control) algorithm and a human forecaster. This study also provided support for the benefit of learning evidence in a novel domain (online dating), serving as a robustness check. In sum, both of our interventions were effective at moving the trust needle. Study 5B examined only that subtle naming manipulation of perceived algorithm learning in a new domain, art, and replicated our core effect (1) as contrasted against an algorithm simply called an algorithm, (2) using a different operationalization of incentive compatibility, and (3) when the human contending with the algorithm was the consumer themself. Finally, Study 6 remained in the realm of art and provided evidence spanning different robustness checks, including lower performance, slower rates of learning, and the stakes of the choice, using a new means by which to make the choice between human and algorithm incentive-compatible (possibly receiving a product resulting from the choice participants made).

## Theoretical and managerial implications

This research takes root in a longstanding and growing tradition that considers how consumers weigh their options in choosing between human and algorithmic forecasters (Dawes, 1979; Grove & Meehl, 1996; Meehl, 1954). To date, this literature has largely specified the circumstances under which consumers tend toward algorithm aversion, eschewing the advice of artificial intelligence (Dietvorst et al., 2015) or, instead, toward algorithm appreciation that adopts it as a source of advice over its human counterpart (Logg et al., 2019). Answers to the question of aversion versus appreciation vary as a function of the domain in which the forecast is being made (Longoni et al., 2019; Longoni & Cian, 2022) and, importantly, these descriptive insights have been utilized in the development of selected interventions that prove capable of tipping the scales one way or another. For instance, Castelo et al. (2019), after finding that consumers especially avoid algorithms for subjective prediction domains, subsequently framed the algorithm in more humanlike ways in order to increase uptake of its advice (see also Schroeder & Epley, 2016 for a successful intervention that incorporates humanness into algorithmic forecasting). Still, when algorithms are portrayed in a manner more like humans, what about this humanlike appearance confers benefits to the algorithm in terms of choice share?

The present investigation answers this question, in part, by incorporating the literature on mistake-making. Prior work had suggested that algorithms suffer because consumers doubt their ability to learn (Dietvorst et al., 2015), meaning that a single mistake reads as diagnostic of an underlying fatal flaw. However, as documented in the heretofore-separate literature on human mistake-making, people who err are seen as capable of learning from those mistakes, and those who deliver on this potential are often held in regard just as high—if not higher—than their non-erring counterparts (Kupor et al., 2018; Reich & Maglio, 2020; see also Reich et al., 2021a). Our synthesis of these two literatures opened the door to the possibility that, should algorithms be framed in a manner conducive to being seen as capable of learning, then similarly providing evidence of past learning should dispel any reservations consumers have in following their advice. Study 2 confirmed that consumers tend to fear that a mistaken algorithm is a broken algorithm incapable of learning, and Studies 3–6 followed from this insight to develop two novel interventions that correct this misperception.

From a theoretical perspective, these findings both deepen and broaden the collective understanding of how consumers think about algorithms. Consistent with prior research on algorithm aversion, we replicate the general tendency of consumers to opt for human forecasts. Yet, in a departure from considerations as to *when* algorithm aversion proves more or less likely, we sought to study *why* it occurs in the first place. Accordingly, our theoretical development did not target particular domains as a moderator of algorithm aversion on the premise that algorithms simply lack the ability to perform in particular domains. Instead, we considered the fundamental question about what consumers believe it means to be a human versus an algorithm. As such, we contribute to a growing range of differences in the perception and evaluation of humans and algorithms, including agency and experience (Epley et al., 2007), motivation (Epley et al., 2008), human features (e.g., speech or voice; Schroeder & Epley, 2016), mind perception (Gray et al., 2007), and factors fundamental to human nature (e.g., emotion, intuition, spontaneity, or soul; Haslam, 2006; Turkle, 2005). While these source features provide avenues for future research to examine differences in how learning is inferred in humans and non-human (e.g., algorithmic) sources, our present results confirmed that consumers innately see humans as dynamic and algorithms as static. If the latter could be conceptualized more like the former, then, our logic proceeded, algorithms might benefit from a similar perceived ability to learn from mistakes. Our data supported these hypotheses, finding that public perceptions of algorithms are malleable. When that malleability manifests as learning from mistakes, as it did in our two interventions, algorithm aversion wanes. In addition to these contributions to theories of human-algorithm differences, our work extends the benefits of mistake-making beyond mistakes made by humans, the primary focus of research to date on this topic. Algorithms, much like humans, garner greater trust when they use their past mistakes as opportunities for growth, raising new possibilities for the literature on the benefits of making, admitting, and learning from mistakes (Reich et al., 2018; Reich & Maglio, 2020).

Whereas our primary theoretical contributions came from documenting how consumers tend to think about humans versus algorithms, our primary managerial implications lie in the theory-derived interventions that we designed and tested. We did not seek to simply showcase that consumers can update their beliefs about algorithms. Rather, we sought to utilize this ability to update in a particular arena: Algorithms can change, to be sure, but our focal type of change is improvement over time that reflects learning from past mistakes. From this broad conceptual vantage point, we designed and tested two novel interventions on the same theme. Studies 3, 4, and 6 presented the performance history of the algorithm that improved over time, an intervention that we then again tested alongside another in Study 5A: a mere change to the name of the algorithmic predicting agent (as well as in Study 5B). Consistently, both of our interventions proved capable of steering consumers toward the belief that algorithms can learn from mistakes and, when they did, they won trust, support, and choice (even in the incentive-compatible designs of Studies 4, 5B, and 6).

Both interventions offer ready-made opportunities for managers to implement in order to increase consumer reliance on the advice provided by their algorithms. While the improving performance histories in Studies 3, 4, 5A, and 6 might only apply to algorithms that have, in fact, improved over time, our data point to two other possibilities of broader relevance. First, let us return to the unexpected finding in Study 4, whereby mention of any performance history led to a nearly even split between preference for a human versus an algorithm. This choice share was substantially higher than that for the algorithm without mention of performance history, suggesting that perhaps mention of any past performance leads consumers to consider that the algorithm has improved over its lifespan in a manner deserving of reasonable trust. However, it is important to note that this result may also be idiosyncratic to the particular performance history used (an 80% success rate), so, we are cautious in speculating about its broad applicability (e.g., below a 50% success rate). More confidence is warranted in our predicted effect from Study 5A and 5B. Second, just about all algorithms in the current marketplace are constantly updating to provide better, more accurate forecasts, matches, and other forms of advice. Accordingly, all of these algorithms can justifiably be called machine learning algorithms, and Study 5A and 5B suggest that those extra words in the name are well worth including in the interest of winning greater consumer trust. Finally, the results from Study 6 especially suggest that any capacity for learning—regardless of levels of performance or rate of improvement—similarly warrant inclusion in promotional materials, as even lower performance and slower improvement still appear sufficient in order to signal the capacity for learning from mistakes and the increase in choice share that results therefrom. This effect, per the null effect of stakes in Study 6, appears robust regardless of whether the stakes of an error are high or low, making the effect documented here relevant to marketing managers for products of both high and low cost and frequent and rare purchase occurrence.

## Limitations and future directions

For our second intervention, designed around the ability of names to change consumer thought and action, we chose "machine learning algorithm" in keeping with the parlance of the times in which our research was conducted. We predicted that, with this common phrase, an easy opportunity to reduce algorithm aversion might be hiding in plain sight. Our results supported this hypothesis, yet it did not probe the necessary and sufficient components in those three words. Specifically, would an algorithm win just as much consumer favor were it simply termed a "learning algorithm?" From a theoretical perspective, any signal of learning from mistakes should prove effective. However, it remains possible that the familiarity of that phrase contributed to its efficacy. To the list that includes algorithms and machine learning, we await future consideration of a third phrase of equal prominence: artificial intelligence. Should the second word in that phrase insinuate that the agent under consideration is capable of using mistakes to improve upon its intelligence, the use of this moniker might offer advantages similar to "machine learning."

After documenting the consumer tendency to doubt the ability of algorithms to learn (from mistakes) in Studies 1 and 2, Studies 3–6 capitalized on this tendency by utilizing novel interventions. These interventions were designed with simplicity as the goal, providing a basic and straightforward test of our theoretical proposition that would also be easy to implement. We made this choice because, as algorithms continue to improve their forecasting performance and further their dominance in most prediction domains, understanding which basic factors cause non-human forecasters to be preferred will be of great value to marketers and consumer behavior researchers alike and could help to improve the uptake of more accurate information (Dawes, 1979; Hastie & Dawes, 2009). In so doing, however, our research cannot speak to where, exactly, consumers believe the learning is taking place. Is all learning from mistakes created equal, or were participants in our studies inferring a particular kind of learning most conducive to winning trust and choice share? Answers to these questions may likely intersect with another important factor in the landscape of algorithm appreciation versus aversion: individual differences. We conjecture that some consumers may inherently trust algorithms while others may refuse to trust algorithms at all. For these types of individuals, the theory-derived interventions identified in the present investigation may be, respectively, unnecessary or ineffective; it could be only those consumers somewhere in the middle that prove most responsive to nudges like ours that increase trust in algorithms. Though outside the scope of the present investigation, determinants of learning specific to algorithms (e.g., computational power, data type, amount of data, or statistical processing method) and individual-level dispositions toward or away from algorithms (e.g., experience, openness to change, technological adoption, education level) could be explored in future research.

As prediction technology continues to improve and broaden the scope of its influence, firms and consumers alike will benefit from better understanding how consumers respond to these new technologies. As society moves from primarily human-based recommendation systems to ones that are predominantly machine-based, research in consumer behavior must continue to work toward an improved understanding of how we think about machines that think. At the same time, the literature on algorithm aversion (and appreciation) will continue to refine the definition of these terms in the first place. Does algorithm aversion always amount to human appreciation, or might algorithm aversion result in simply avoiding making a choice in which the algorithm provides a recommendation while advice from a human is unavailable? With this research, we identify that consumers often think that machines cannot think—at least not to the extent that thinking is learning and learning is thinking. By identifying this tendency, interventions like the two designed and tested here can give rise to new opportunities by which to disabuse consumers of the notion that machines cannot learn from mistakes, in turn fostering greater reliance on algorithms and, via their tendency to outperform humans, to foster greater satisfaction and wellbeing.

## CONFLICT OF INTEREST

We have no known conflict of interest to disclose.

## ORCID

*Taly Reich* https://orcid.org/0000-0002-2978-4406
*Sam J. Maglio* https://orcid.org/0000-0003-0691-1721

## REFERENCES

Agrawal, A., Gans, J., & Goldfarb, A. (2018). *Prediction machines: The simple economics of artificial intelligence*. Harvard Business Press.

Arkes, H. R., Dawes, R. M., & Christensen, C. (1986). Factors influencing the use of a decision rule in a probabilistic task. *Organizational Behavior and Human Decision Processes*, *37*(1), 93–110.

Berger, B., Adam, M., Rühr, A., & Benlian, A. (2021). Watch me improve—Algorithm aversion and demonstrating the ability to learn. *Business & Information Systems Engineering*, *63*(1), 55–68.

Berlo, D. K., Lemert, J. B., & Mertz, R. J. (1969). Dimensions for evaluating the acceptability of message sources. *Public Opinion Quarterly*, *33*(4), 563–576.

Binzel, C., & Fehr, D. (2013). Social distance and trust: Experimental evidence from a slum in cairo. *Journal of Development Economics*, *103*, 99–106.

Bolger, F., & Wright, G. (1992). Reliability and validity in expert judgment. In *Expertise and decision support* (pp. 47–76). Springer.

Brockman, J. (2015). *What to think of machines that think*. Harper Perennial.

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

Chesney, T., & Su, D. (2010). The impact of anonymity on weblog credibility. *International Journal of Human-Computer Studies*, *68*(10), 710–718.

Dana, J., & Thomas, R. (2006). In defense of clinical judgment and mechanical prediction. *Journal of Behavioral Decision Making*, *19*(5), 413–428.

Dawes, R. M. (1979). The Robust Beauty of Improper Linear Models in Decision Making. *American Psychologist*, *34*(7), 571–582.

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science*, *243*, 1668–1674.

Dietvorst, B. J., & Bharti, S. (2020). People reject algorithms in uncertain decision domains because they have diminishing sensitivity to forecasting error. *Psychological Science*, *31*(10), 1302–1314.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneouslyavoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, *144*(1), 114–126.

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.

Dijkstra, J. J. (1999). User Agreement With Incorrect Expert System Advice. *Behaviour & Information Technology*, *18*(6), 399–411.

Dijkstra, J. J., Liebrand, W. B. G., & Timminga, E. (1998). Persuasiveness of expert systems. *Behaviour & Information Technology*, *17*(3), 155–163.

Dweck, C. (2011). Implicit theories. In *Handbook of theories in social psychology* (pp. 43–61). Sage Publications.

Eastwood, J., Snook, B., & Luther, K. (2015). Measuring the reading complexity and oral comprehension of canadian youth waiver forms. *Crime & Delinquency*, *61*(6), 798–828.

Edmondson, A. C. (1996). Learning from mistakes is easier said than done: Group and organizational influences on the detection and correction of human error. *Journal of Applied Behavioral Science*, *32*(1), 5–28.

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review*, *114*(4), 864–886.

Epley, N., Waytz, A., Akalis, S., & Cacioppo, J. T. (2008). When we need a human: Motivational determinants of anthropomorphism. *Social Cognition*, *26*(2), 143–155.

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science*, *315*(5812), 619.

Grove, W. M., & Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical–statistical controversy. *Psychology, Public Policy, and Law*, *2*(2), 293–323.

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment*, *12*(1), 19–30.

Haslam, N. (2006). Dehumanization: An integrative review. *Personality and Social Psychology Review*, *10*(3), 252–264.

Hastie, R., & Dawes, R. M. (2009). *Rational choice in an uncertain world: The psychology of judgment and decision making*. Sage Publications.

Hayes, A. F. (2013). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. Guilford Press.

Highhouse, S. (2008). Stubborn reliance on intuition and subjectivity in employee selection. *Industrial and Organizational Psychology*, *1*(3), 333–342.

Hong, Y.-y., Chiu, C.-y., & Dweck, C. S. (1995). Implicit Theories of Intelligence. In *Efficacy, agency,and self-esteem* (pp. 197–216). Springer.

Hovland, C. I., Janis, I. L., & Kelley, H. H. (1953). *Communication and persuasion*. Yale University Press.

Hovland, C. I., & Weiss, W. (1951). The influence of source credibility on communication effectiveness. *Public Opinion Quarterly*, *15*(4), 635–650.

Kunda, Z. (1999). *Social Cognition: Making Sense of People*. MIT Press.

Kupor, D., & Laurin, K. (2020). Probable cause: The influence of prior probabilities on forecasts and perceptions of magnitude. *Journal of Consumer Research*, *46*(5), 833–852.

Kupor, D., Reich, T., & Laurin, K. (2018). The (bounded) benefits of correction: The unanticipated interpersonal advantages of making and correcting mistakes. *Organizational Behavior and Human Decision Processes*, *149*, 165–178.

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.

Longoni, C., Bonezzi, A., & Morewedge, C. K. (2019). Resistance to medical artificial intelligence. *Journal of Consumer Research*, *46*(4), 629–650.

Longoni, C., & Cian, L. (2022). Artificial intelligence in utilitarian vs. hedonic contexts: the "word-of-machine" effect. *Journal of Marketing*, *86*(1), 91–108.

Maglio, S. J., & Feder, M. A. (2017). The making of social experience from the sounds in names. *Social Cognition*, *35*, 663–674.

Maglio, S. J., & Polman, E. (2014). Spatial orientation shrinks and expands psychological distance. *Psychological Science*, *25*, 1345–1352.

Maglio, S. J., & Polman, E. (2016). Revising probability estimates: Why increasing likelihood means increasing impact. *Journal of Personality and Social Psychology*, *111*(2), 141–158.

Maglio, S. J., Rabaglia, C. D., Feder, M. A., Krehm, M., & Trope, Y. (2014). Vowel sounds in words affect mental construal and shift preferences for targets. *Journal of Experimental Psychology: General*, *143*(3), 1082–1096.

McGuire, W. J. (1978). An information-processing model of advertising effectiveness. *Behavioral and Management Science in Marketing*, *15*, 156–180.

Meehl, P. E. (1954). *Clinical versus statistical prediction: a theoretical analysis and a review of the evidence*. University of Minnesota Press.

Mitchell, T. M. (1997). *Machine learning*. McGraw-Hill.

Önkal, D., Goodwin, P., Thomson, M., Gönül, S., & Pollock, A. (2009). The relative influence of advice from human experts and statistical methods on forecast adjustments. *Journal of Behavioral Decision Making*, *22*(4), 390–409.

Palmer, M., Simmons, G., & de Kervenoael, R. (2010). Brilliant mistake! Essays on incidents of management mistakes and mea culpa. *International Journal of Retail & Distribution Management*, *38*(3), 234–257.

Parker, D., & Lawton, R. (2003). Psychological contribution to the understanding of adverse events in health care. *Quality and Safety in Health Care*, *12*(6), 453–457.

Pornpitakpan, C. (2004). The Persuasiveness of source credibility: A critical review of five decades' evidence. *Journal of Applied Social Psychology*, *34*(2), 243–281.

Promberger, M., & Baron, J. (2006). Do patients trust computers? *Journal of Behavioral Decision Making*, *19*(5), 455–468.

Rabaglia, C. D., Maglio, S. J., Krehm, M., Seok, J. H., & Trope, Y. (2016). The sound of distance. *Cognition*, *152*, 141–149.

Reich, T., Kupor, D. M., & Smith, R. K. (2018). Made by mistake: When mistakes increase product preference. *Journal of Consumer Research*, *44*(5), 1085–1103.

Reich, T., & Maglio, S. J. (2020). Featuring mistakes: The persuasive impact of purchase mistakes in online reviews. *Journal of Marketing*, *84*(1), 52–65.

Reich, T., Maglio, S. J., & Fulmer, A. G. (2021a). No laughing matter: Why humor mistakes are more damaging for men than women. *Journal of Experimental Social Psychology*, *96*, 104169.

Reich, T., Savary, J., & Kupor, D. (2021b). Evolving choice sets: The effect of dynamic (vs. static) choice sets on preferences. *Organizational Behavior and Human Decision Processes*, *164*, 147–157.

Reich, T., & Tormala, Z. (2013). When contradictions foster persuasion: An attributional perspective. *Journal of Experimental Social Psychology*, *49*(3), 426–439.

Schroeder, J., & Epley, N. (2016). Mistaking minds and machines: How speech affects dehumanization and anthropomorphism. *Journal of Experimental Psychology*, *145*, 1427–1437.

Shaffer, V. A., Adam Probst, C., Merkle, E. C., Arkes, H. R., & Medow, M. A. (2013). Why do patients derogate physicians who use a computer-based diagnostic support system? *Medical Decision Making*, *33*(1), 108–118.

Shrum, L. J., & Lowrey, T. M. (2007). Sounds convey meaning: The implications of phonetic symbolism for brand name construction. In *Psycholinguistic Phenomena in Marketing Communications* (pp. 39–58). Erlbaum.

Stefaniak, C., & Robertson, J. C. (2010). When auditors err: How mistake significance and superiors' historical reactions influence auditors' likelihood to admit a mistake. *International Journal of Auditing*, *14*(3), 41–55.

Sternthal, B., Phillips, L. W., & Dholakia, R. (1978). The persuasive effect of scarce credibility: A situational analysis. *Public Opinion Quarterly*, *42*(3), 285–314.

Tjosvold, D., Zi-you, Y., & Hui, C. (2004). Team learning from mistakes: The contribution of cooperative goals and problem-solving. *Journal of Management Studies*, *41*(7), 1223–1245.

Turkle, S. (2005). *The second self: Computers and the human spirit*. Mit Press.

Uribe, C. L., Schweikhart, S. B., Pathak, D. S., Marsh, G. B., & Reed Fraley, R. (2002). Perceived barriers to medical-error reporting: An exploratory investigation. *Journal of Healthcare Management*, *47*(4), 263–279.

Yeomans, M., Shah, A., Mullainathan, S., & Kleinberg, J. (2019). Making sense of recommendations. *Journal of Behavioral Decision Making*, *32*(4), 403–414.

Yorkston, E., & Menon, G. (2004). A sound idea: Phonetic effects of brand names on consumer judgments. *Journal of Consumer Research*, *31*(1), 43–51.

**How to cite this article:** Reich, T., Kaju, A., & Maglio, S. J. (2022). How to overcome algorithm aversion: Learning from mistakes. *Journal of Consumer Psychology*, *00*, 1–18. https://doi.org/10.1002/jcpy.1313