Beyond Algorithm Aversion: The Impact of Conventionality on Evaluation of Algorithmic and Human-Made Errors

by

Hamza Tariq

A thesis

presented to the University of Waterloo

in fulfilment of the

thesis requirement for the degree of

Master of Arts

in

Psychology

Waterloo, Ontario, Canada, 2024

© Hamza Tariq 2024

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

I developed the study design in collaboration with Dr. Jonathan Fugelsang and Dr. Derek Koehler. Additionally, I conducted the testing, data collection, and data analysis for all studies. All drafts of the thesis were primarily written by myself, with substantial input from Dr. Jonathan Fugelsang and Dr. Derek Koehler.

Abstract

Prior research has found that when an algorithm makes an error, people judge it more severely than when the same mistake is made by a human. This bias, known as algorithm aversion, was investigated across two studies (N = 1199). Specifically, we explored the effect of the status quo on people's reactions to identical mistakes made by humans and algorithms. We found significant algorithm aversion when participants were informed that the decisions described in the scenarios are conventionally made by humans. However, when participants were told that the same decisions are conventionally made by algorithms, the bias diminishes, is eliminated, or even reverses direction. This effect of varying whether the algorithm or the human is described as the convention had a particularly strong influence on recommendations of which decision maker should be used in the future. These findings suggest that the existing status quo has a consequential influence on people's judgments of mistakes. Implications for people's evolving relationship with algorithms and technology are discussed.

Acknowledgements

I would like to express my deepest gratitude to my supervisors, Dr. Jonathan Fugelsang and Dr. Derek Koehler, for their exceptional patience and unwavering support. They gave me the open space and freedom to explore my own curiosities, always providing feedback that made me think of things I never thought of before, and helped shape my musings into ideas worthy of scientific exploration.

There are a number of people whom I would like to credit with helping me on the journey of this research project over the past two years. These include Alexandra van der Valk, Kaiden Stewart, Ethan Meyers, Kristina Wu, Dogukan Demircioglu, and Xinyi Lu. I thank them for helping me with the tools and practical skills I needed to complete this work, the regular feedback and discussion sessions, guiding me through two years of graduate school, being mentors, and most importantly, being the most encouraging and understanding of friends. Then, a big thank you to everyone who has been a part of our lab in the past two years and the cognitive area, who, both in and outside of our weekly seminars, have been the most helpful and the best of colleagues anyone could ask for.

Finally, I would like to thank my supervisors, Dr. Evan Risko, Dr. Sam Johnson, and everyone at the psychology department whose extraordinary personal support behind the scenes this past year, in the most challenging of times, helped ensure that I could bring this work and my degree to their completion.

Dedication

This thesis is dedicated to my partner, Sarah Anwer. Words cannot express my gratitude for your love and patience. Thank you for everything. A special dedication to our dogs, Bella and Maddie, whose contributions in therapeutic value will forever remain unmatched.

Table of Contents

Author's Declaration ii
Statement of Contributions iii
Abstract iv
Acknowledgements v
Dedication vi
List of Tables and Figures viii
1. Introduction 1
Overview of Studies. 5
2. Study 1 7
2.1. Method 7
2.1.1. Participants 7
2.1.2. Procedure 7
2.2. Results 11
2.2.1. Mistake Severity & Concern Measure. 11
2.2.2. Retention Measure. 13
2.2.3. Recommendation for the Future. 15
2.2.4. Summary of Findings 16
3. Study 2 18
3.1. Method 18
3.1.1. Participants 18
3.1.2. Procedure 18
3.2. Results 20
3.2.1. Mistake Severity & Concern Measure. 20
3.2.2. Retention Measure. 22
3.2.3. Recommendation for the Future. 24
3.2.4. Summary of Findings 25
4. General Discussion 26
References 32

List of Tables and Figures

Table 1 ..... Page 11
Figure 1 ..... Page 8
Figure 2 ..... Page 12
Figure 3 ..... Page 14
Figure 4 ..... Page 15
Figure 5 ..... Page 19
Figure 6 ..... Page 21
Figure 7 .....Page 23
Figure 8 ..... Page 24

1. Introduction

1) Everything that's already in the world when you're born is just normal;

2) anything that gets invented between then and before you turn thirty is incredibly exciting and creative and with any luck you can make a career out of it;

3) anything that gets invented after you're thirty is against the natural order of things and the beginning of the end of civilization as we know it until it's been around for about ten years when it gradually turns out to be alright really.

- Douglas Adams (1999), *How to Stop Worrying and Learn to Love the Internet*

Throughout human history, the 'normal' or conventional ways of doing things have always evolved with the advent of new inventions. Manual scribing gave way to the printing press, magnetic compasses replaced celestial navigation, steam engines revolutionized how we built and traveled, and pocket calculators transformed complex manual calculations. What might have initially been perceived as "against the natural order of things" almost always—over time—turned out, as Douglas Adams put it 25 years ago, to be "alright really." It is a fair assumption that this change of conventions has accelerated in the more recent past. Computer algorithms, and increasingly AI, have become deeply embedded in the fabric of contemporary life. From daily navigation, learning, and job searching, to socializing, dating and entertainment, algorithm-based applications—such as Google Maps, LinkedIn, Bumble, Tiktok, or Netflix—feature prominently. Large language models (LLMs) —like OpenAI's ChatGPT, Meta's Llama, or Google's Bard—and other AI systems are in a race to fundamentally change how we work, educate, live, and interact with our world. The present era has been quite aptly described as the "Age of Algorithms" (Alang, 2016). The rapid saturation and turnover of algorithm-driven technologies mean that the examples just listed above may be current and relevant at the time of writing, but by the time one reads this, they may already have been heaped onto the ever-growing pile of technological obsolescence.

The term "algorithm" has evolved in its usage and connotation in recent times. The word itself comes from a Latinized version of the name of a 9th-century Persian polymath, al-Khwārizmī, who is credited with foundational contributions to algebra, arithmetic, astronomy, and geography (Britannica, 2024). Historically, "algorithm" has referred to a set of well-defined instructions, a step-by-step procedure for solving mathematical problems, a series of

calculations which compute a function, or even an evidence-based forecasting formula (Dietvorst et al., 2015; Logg et al., 2019; Rogers, 1987). Today, the term "algorithm" may, in common parlance, refer to various modern technologies—from search engines and social media feeds to several AI-driven applications (Raine & Anderson, 2017). All these definitions of and references to algorithms converge on a fundamental reality—that they are all sets of instructions dedicated to specific tasks. In this paper, to cover both historical definitions and contemporary uses, we use the word 'algorithm' to refer to any automated set of instructions designed to perform calculations for a specific function.

Unlike other tools that were embraced into daily use, algorithms have not been universally welcomed in human decision-making, despite their proven track record of superior judgment, forecasting, and reliability compared to human agents. For example, algorithms have been consistently found to be more accurate than human experts across several domains including medicine, psychology, criminal justice, and academic performance (Dawes, 1979; Dawes et al., 1989; Meehl, 1954). Instead, what has often been observed is a human distrust of algorithms—now labeled algorithm aversion—which has been consistently reported as far back as the 1950s to the present (Dietvorst et al., 2015; Eastwood et al., 2012; Meehl, 1954, Promberger and Baron, 2006; Önkal et al., 2009, Shaffer et al., 2013). It seems plausible that humans would not trust algorithms as much as other humans in various contexts. However, empirical research on people's perceptions of algorithms and their responses to algorithmic advice has been relatively sparse until the recent past (Logg, 2022). It been speculated that algorithm aversion may be a byproduct of human expectations and assumptions such as: algorithms should be near perfect and free of biases; algorithms are unable to learn from experience unlike human decision-makers; algorithms would continue to make systematic errors; algorithms may not consider important qualitative information or unique circumstances; many modern algorithmic systems are black boxes; or algorithms might be dehumanizing or unethical to use for important decisions (Bigman & Grey, 2018; Bonezzi et al., 2021; Castelo et al., 2019; Dawes, 1979; Einhorn, 1986; Grooves & Meehl, 1996; Highhouse, 2008, Longoni et al., 2019).

As detailed above, various reasons might explain people's bias against algorithms compared to human agents. This bias becomes more apparent in situations where humans and algorithms make identical mistakes—but people perceive the errors made by algorithms as worse and penalize them more severely. In a series of five studies, Dietvorst et al. (2015) asked participants to predict outcomes from real data, such as MBA students' success and the number of airline passengers departing from U.S. states. Participants had to choose whether to rely on their own forecasts, another human's forecasts, or an algorithm's forecasts. The results consistently showed that participants who observed the algorithm making forecasts were less likely to choose it, even if the algorithm outperformed human forecasters. These findings revealed that seeing an algorithm err reduced participants' confidence in the algorithm's forecasts more significantly than seeing a human err reduced their confidence in the human's forecasts. In other words, people exhibited greater intolerance for algorithmic errors. This effect—algorithm aversion—was observed across different forecasting domains and persisted even when the algorithm's overall performance was superior to human forecasters. The authors concluded that algorithm aversion is at least partly driven by people's experience with the algorithm—particularly when they witness it making mistakes. One can infer from these studies that people's aversion to

algorithms is not entirely rational, as it leads to a preference for inferior human forecasters even when better decision aids are available. This raises the question: why is there greater intolerance for mistakes made by algorithms when, in principle, an identical error should be considered equally problematic regardless of who makes it? Is this bias purely stemming from the perceived impersonal, blackbox, and non-human nature of algorithms and the high expectations set for their performance, as discussed earlier, making any error seem more significant? Moreover, will this aversion persist in the future, and under what circumstances might it change?

Building on this key work on algorithm aversion in the context of mistakes (Dietvorst et al., 2015), our research sought to examine whether people always exhibit a higher bias against algorithms or if the bias is also affected by which option—human or algorithm—is perceived as the *convention* in that domain. As the context of our interactions with algorithms evolves, many of the stated reasons for algorithm aversion might become less relevant. In this research, we focused on one such contextual factor—the perceived conventional or non-conventional status of the error maker. Our interpretation of 'conventionality' is influenced by aspects of status quo bias (Samuelson & Zeckhauser, 1988), omission bias (Ritov & Baron, 1992), default options in choice architecture (Johnson & Goldstein, 2003; Johnson et al., 2011; Thaler & Sunstein, 2008), endowment effects and loss aversion (Dinner et al, 2014; Kahneman et al., 1991; Thaler, 1980), and other works in this broad area of research. By 'conventional,' we refer to the default or status quo options that are trusted and accepted within a given context or domain.

For example, in the context of medical diagnosis, a human doctor's opinion is typically seen as the conventional option, whereas the analysis provided by an algorithmic diagnostic robot might be considered the non-conventional option. Consider a scenario where an experienced and skilled human doctor, and a well-tested and accurate diagnostic robot, both make an identical incorrect diagnosis. Setting aside the inherent differences between human and algorithmic agents discussed earlier, we suggest that the convention status of the error makers also influences how people judge identical mistakes. For instance, if the human doctor is perceived as the conventional choice, their mistake might be more easily forgiven due to people's long history of relying on doctors, awareness of the doctor's extensive training and expertise, and the established trust and status they hold within the medical community. Conversely, the erring algorithmic diagnostic robot, perceived as an alternative, less familiar, and less established tool—might, understandably, draw more scrutiny and skepticism. These more adverse reactions could then result in harsher judgment and penalties for the robot, despite making the same error as the human doctor.

In contrast, in the context of everyday navigation and route planning, algorithmic GPS systems like those in navigation devices (e.g., Garmin) and mobile applications (e.g., Google Maps, Waze) can be presumed to be the conventional options today (Yang et al., 2020). Conversely, human-given directions or using a paper map might be considered non-conventional navigation options. Consider a situation where a GPS app and a human friend providing directions both make an identical navigational error, leading to a significant delay in the driver reaching the destination. As algorithmic navigation systems are perceived as the conventional option, the app's error might be judged less severely. This might be due to the driver's historic familiarity with the app's everyday accuracy, the perception of the error as more

of an isolated glitch than a systematic issue, and the app's trusted and well-established role in navigation reflected in its widespread use. On the other hand, the human friend might be judged more critically for the same error. This might be because, in today's world, asking a person for directions is less common and might be perceived as a nonessential alternative for what has become a highly automated task. Moreover, due to less familiarity with the friend's general navigation skills, the driver might be quicker to judge the mistake as indicative of poor navigation ability—making them lose confidence in the friend's ability to give directions in the future.

Building on the context of conventionality, we developed the working concept of *alternate aversion*. We propose that alternate aversion occurs when the presence of a conventional option leads people to have a stronger aversion to the non-conventional or alternate option. Consequently, in the aftermath of an identical mistake, people will show a stronger aversion to the non-conventional option by judging and penalizing its error more severely than they do the conventional option. This suggests that in the context of algorithm aversion, perhaps it is not just the *algorithm* itself which drives the aversion—but its typical status as the non-conventional option in the human-algorithm comparison which contributes to its harsher treatment.

We posit that, psychologically, such an aversion could be explained by "conventional" factors. For instance, status quo bias suggests that people prefer familiar and established options, perceiving them as safer and more reliable—which could be reflected as a preference for conventional 'human' options in human-algorithm comparisons. Omission bias suggests that people are more forgiving of errors that arise from inaction. Thus, a mistake made by the conventional option might be viewed more leniently since it is a choice that would be made most of the time anyway—similar to an error of inaction. However, choosing a non-conventional option and then seeing it make a mistake might be seen as an error resulting from action and, therefore, less forgivable. Choice architecture research shows presenting options as defaults can influence perceptions and decisions in their favor. Hence, when the conventional option is seen as the default choice, identical errors by the 'non-default' or alternate option could be judged as relatively worse. Additionally, there might be a tendency to overestimate the importance of the conventional option to the entire system and how disruptive it might be to replace it. This bias might be considered a form of 'reference dependence' in the default effects literature (Dinner et al., 2011). The default option acts as a reference point, influencing how alternative or non-default options are evaluated as gains or losses. Dinner and colleagues (2011) suggested that defaults could function as "instant endowments," meaning decision-makers may perceive themselves as having already chosen the default option—and use that as their reference point. This aligns with the endowment effect, where people value an object more once they own it (Thaler, 1980). This effect can increase the perceived value of the default option and activate loss aversion—where the pain of losing something is much more powerful than the pleasure of gaining something of equal value (Tversky & Kahneman, 1991). The bias in the evaluation of conventional and alternate options may result from this interplay of reference dependence and loss aversion (Kahneman et al., 1991).

As with the medical diagnosis example above, many comparisons of human-algorithm judgments are often made in contexts where the human is perceived as the conventional or

default option for the task at hand. We believe that this inherent bias can skew the evaluation of errors, making it crucial to understand how the context of conventionality affects judgment. In this paper, we will explore this by experimentally manipulating whether humans or algorithms are described as the conventional option for the same task. By doing so, we hope to examine if and how the status of being the "conventional" or "alternate" option influences the severity of judgment when mistakes are made, and how it impacts future decisions regarding the error makers.

Overview of Studies.

For both studies in this paper, we hypothesized that in comparative judgments of humans and algorithms, biases in judgments of identical errors would be influenced by the conventionality status of the error maker. Specifically, we predicted that (1) errors made by non-conventional decision-makers would be perceived as more severe and concerning, and (2) despite making identical mistakes, conventional decision-makers would be retained and recommended for future tasks more frequently than non-conventional ones. In Study 1, participants observed human and algorithmic decision-makers making errors while screening applicants in a college admissions scenario. In Study 2, participants saw both types of decision-makers making errors while evaluating the quality of sound speakers in a product quality control scenario. Across both experiments, participants' evaluations of errors differed depending on whether the error maker—human or algorithm—was presented as the conventional or non-conventional option. We also demonstrated that the conventionality of the error maker affected participants' decisions on future retention and recommendations of the error making agent, with conventional options often being favored more frequently.

Setting up Conventionality. In both of the following experiments, the so-called conventional, status quo or default option of the decision maker was set up using three characteristics: historic use, prevalence, and perceived system dependence.

Historic use. The conventional decision-maker — human or algorithmic — was described as having been in the role for the past 10 years. This history may carry potential implications: a long tenure might indicate familiarity, stability, reliability, and effectiveness over time. Similar to the status quo bias (Samuelson & Zeckhauser, 1988), it might suggest a pre-existing familiarity and trust in the decision-maker, in contrast to the unfamiliarity and uncertainty associated with a newer decision-maker who lacks a historical track record.

Prevalence. The conventional decision-making option was described as the most commonly used and widespread approach in similar contexts, such that 85% of operators or the industry used the conventional option. We assume that people are likely to trust existing methods that are widely recognized and adopted as standard practice. As such, widespread use might indicate general social acceptance and collective endorsement of the decision-maker by a majority. This reinforces its position as the conventional method and stands in contrast to an alternative option — new or old — which is not in common use.

System Dependence. Alongside its historic and prevalent use, the conventional method was described as crucial to the continued functioning of the system of which it is part. The conventional method is seen as a default reference point which makes people averse to penalizing it, while an alternate option is perceived as relatively more expendable. Subsequently, harsh penalties for the conventional option might be perceived as a threat to the overall function, but there is little threat to function if the alternate option is penalized. To elaborate on system dependence, consider this example: the system of road transportation is dependent on human-driven cars at this point in time.

Human-driven cars are the conventional method due to their widespread use and their central place in the current infrastructure. They are the primary means of private and commercial transport around the world. If a new, alternate method—such as autonomous vehicles—is introduced, any errors or accidents associated with these autonomous vehicles might be judged more harshly. This is because the autonomous vehicles are seen as non-essential to the system's current functioning and can be penalized without significantly disrupting the overall transportation system. Conversely, errors made by human-driven cars might be more readily overlooked or forgiven, as penalizing them too harshly could be seen as destabilizing the entire transportation infrastructure.

2. Study 1

In Study 1, we created a scenario where both human and algorithmic decision-makers made errors while screening college applicants, focusing specifically on prediction-based errors. These errors occurred when either human admissions officers or admissions algorithms failed to accurately predict future outcomes (college performance) based on available information (the admissions application). We utilized the three "convention" characteristics—history, prevalence, and system dependence—to establish which decision-making agent was the conventional option and which was the alternate option in each condition (Figure 1).

2.1. Method

Data and materials for this experiment is available on the Open Science Framework (OSF) here: https://osf.io/29sj7/.

2.1.1. Participants

We recruited participants for this study (N = 594, 308 women; 279 men; 6 non-binary; 1 other; $Mdn_{\text{age group}}$ = 35-44) through CloudResearch's MTurk Toolkit. Participants were recruited simultaneously for both studies in this paper but could only participate in one. The recruitment criteria included: U.S. residency; completion of at least 100 prior HITs; a HIT approval rating of 99% or higher; and no participation in an earlier pilot study of this project. All participants responded to two comprehension checks before starting the study.[1]

2.1.2. Procedure

Participants were instructed to review a case study before being asked to make their judgments. The fictitious case study in this experiment involved college admissions. All participants were briefed on the scenario where a state higher education commission was evaluating admissions practices. The review focused on the performance of human admission officers versus computerized admissions algorithms in processing applications.

Participants were randomly assigned to one of four conditions in a 2 X 2 between-subjects design. First, participants were introduced to either a human officer or a computer algorithm as the conventional and established method of reviewing admissions applications at a hypothetical college. Next, the commission was shown to have run tests in which errors were made by either the human officer or the algorithm — with the mistake being a failure to predict just over 50% of the students who would not graduate due to poor academic performance. In the scenario, only one of the decision-making systems, either the human or the algorithm, was said

[1] We also included two pre-study comprehension check screening questions. However, due to a programming error, participants were allowed to correct their answers to pre-study comprehension checks even if they initially answered incorrectly.

to have been tested and found to make errors. This factor was varied independently of which was said to be the convention; that is, in some conditions the convention was tested and found to make errors, while in other conditions it was the alternative, non-conventional option that was tested and found to make errors. In summary, the experiment had a two (convention: human vs. algorithm) by two (error maker: human vs. algorithm) between-subject design. The complete information regarding the conventional method and the error, as presented to the participants, is detailed in Figure 1.

Participants completed comprehension and memory checks both after learning about the conventional method and again after the error simulation to ensure they understood and retained the provided details. If participants answered incorrectly, they were given the relevant information again and were required to select the correct answer to proceed further.[2]

Finally, participants were asked to evaluate: (1) the severity of the mistake on a 1 (Not at all serious) to 6 (Extremely serious) scale; (2) their level of concern regarding the continued use of the error-prone method on a 1 (Not all concerned) to 6 (Extremely concerned); (3) whether the fictional college should maintain its conventional admissions approach on a 1 (Definitely not) to 6 (Definitely) scale; and (4) which admissions option—human or algorithm—they believed should be preferred by the state in the future. Table 1 provides details on each of these questions. The experiment concluded with a series of demographic questions.

Figure 1

College Admissions Scenario as Observed by Participants

Part 1
Case Background The higher education commission is currently inspecting college admissions practices across the state. This inspection is part of the state's upcoming five-year strategy for higher education. Specifically, the commission is evaluating two different approaches for assessing undergraduate college applications: a human admissions officer or a computerized

[2] The comprehension checks used in this and the following study are detailed in the Materials documents on the OSF page for this paper (https://osf.io/29sj7/).

algorithm designed to review admissions applications.

Part 2

Case Details

College SY is one of the state colleges being inspected by the commission.

For the past 10 years, [College SY has employed Jordan Taylor, a human admissions officer, to assess applications / College SY has used ScholarVision 3.0, an admissions computer algorithm, to assess applications.]

These days, most colleges (around *85%*) [use an admissions officer, while the rest use a computer program to review their undergraduate applications / use a computer program, while the rest use a human officer to review their undergraduate applications.]

As it can take a lot of time and resources to [develop custom-built admissions computer algorithms, colleges across the state currently depend on human admissions officers, like Jordan Taylor, to keep their admissions processes on track / train qualified human admissions officers, colleges across the country currently depend on admissions-focused computer algorithms, like ScholarVision 3.0, to keep their admissions processes on track.]

Part 3

The higher education commission decides to run multiple tests on both options before making a final recommendation on college admissions practices.

Part 4

Test: Human Admissions Officer

In one test, the commission tasked Jordan Taylor, [the admissions officer / an admissions officer employed elsewhere in the state], to assess the admissions applications of 100 former students of College SY. Out of these, 30 students were repeatedly placed on academic probation and consequently failed to graduate. These 30 were deliberately inserted into the 100 applications reviewed, for testing purposes.

Taylor was then asked to **pick the 30 students from the list of 100** who were unlikely to graduate based on their expected academic performance in college.

When the test results were reviewed, the commission found that Taylor had ***failed to include 16 of the 30 students who could not graduate*** *due to poor academic results.*

OR

Test: Computerized Admissions Algorithm

In one test, the commission tasked ScholarVision 3.0, [an admissions algorithm used elsewhere in the state / the admissions algorithm], to assess the admissions applications of 100 former students of College SY. Out of these, 30 students were repeatedly placed on academic probation and consequently failed to graduate. These 30 were deliberately inserted into the 100 applications reviewed, for testing purposes.

ScholarVision was then asked to **pick the 30 students from the list of 100** who were unlikely to graduate based on their expected academic performance in college.

When the test results were reviewed, the commission found that ScholarVision had **failed to include 16 of the 30 students who could not graduate** due to poor academic results.

Note. This includes all scenario parts of Study 1 in sequence, as seen by the participants, barring the comprehension questions. Colored text within square brackets shows variations made per condition (blue = human convention, red = algorithm convention).

Table 1

All Measures Presented to Participants Following the Error

Measure	Study	Questions	Scale or Options
Mistake Severity	Study 1	According to your judgment, how serious is this error by the [human admissions officer OR admissions algorithm]?	1 (Not serious at all) to 6 (Extremely serious)
	Study 2	According to your judgment, how serious is this error by the [sound-quality analyst OR sound-quality algorithm]?	
Level of Concern	Study 1	How concerned should the commission be, given this mistake, about the use of the [the human officer OR computerized algorithm] in statewide college admissions?	1 (Not at all concerned) to 6 (Extremely concerned)
	Study 2	How concerned should the company be, given this mistake, about the use of the [human analyst OR computerized algorithm] in quality control for its speakers?	
Retention	Study 1	Should College SY's [admissions officer OR computer program], [Jordan Taylor OR ScholarVision 3.0], be retained after the results of this test?	1 (Definitely not) to 6 (Definitely)
	Study 2	Should the company's [analyst OR computer program], [Jordan Taylor OR SonicVerifier 3.0], be retained after the results of this test?	
Future Choice	Study 1	Which college admissions approach do you think the commission should recommend for the state's future five-year higher education plan?	Human admissions officers OR Customized admissions algorithms
	Study 2	Which sound-quality testing method do you think the company should choose for their next five-year cycle?	A human sound-quality analyst OR A sound-quality algorithm

Note. Depending on the condition, each participant in both studies was asked only about the option within square brackets that made the error. The order of choices in the last measure (Future Choice) was randomized for all participants.

2.2. Results

2.2.1. Mistake Severity & Concern Measure.

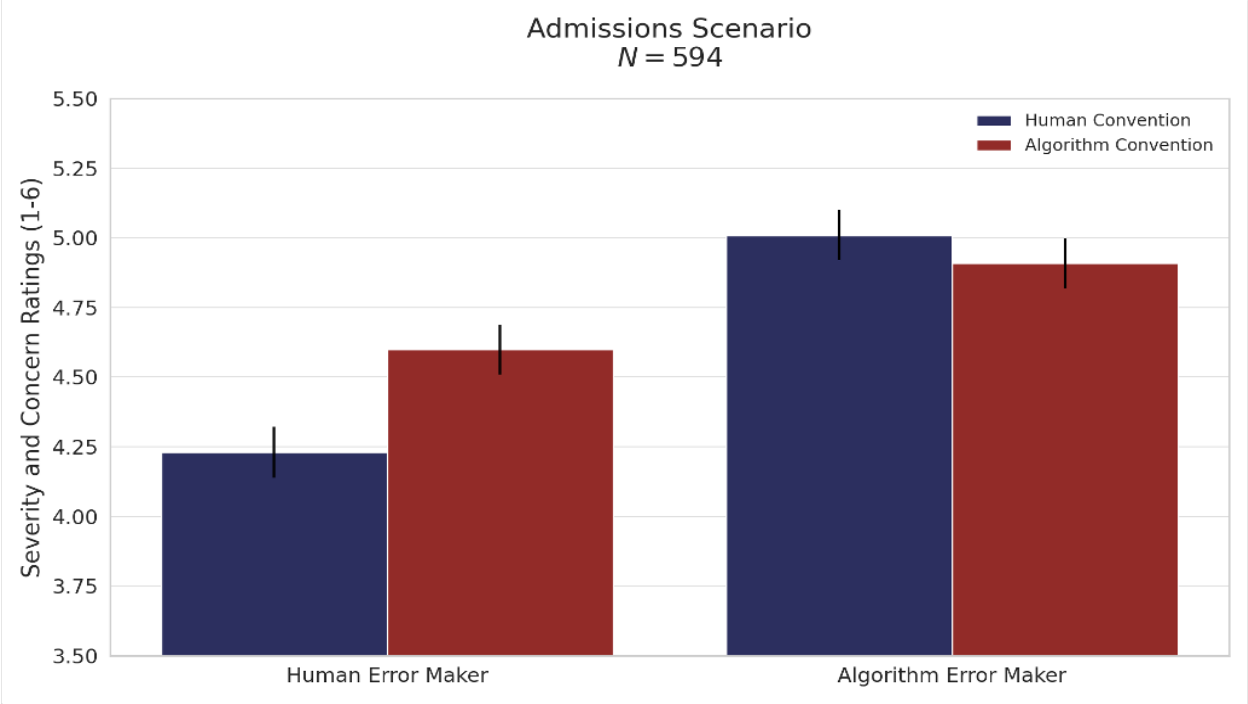
We first conducted a correlation analysis to assess the relationship between the mistake severity and level of concern ratings. There was a strong, positive correlation between the two variables, $r(592) = .798$, $p < .001$, 95% CI [0.767, 0.825]. The high correlation suggested that the variables were capturing similar aspects of participants' judgment — that is, negative reactions to errors — and hence we combined the two to avoid redundancy. This was done by creating a new variable by averaging the two variables across all participants.

We examined participants' responses using a convention (human convention, algorithm convention) by error maker (human admissions officer, computerized admissions algorithm) ANOVA (Figure 2). There was a main effect of the error maker, ($F(1, 590) = 36.34$, $p < .001$, $\eta^2_P = 0.06$), indicating that overall, participants judged the mistakes of the computerized admissions algorithm more severely ($M = 4.96$, $SD = 1.00$) than they did the mistakes of the

human admissions officer (M = 4.42, SD = 1.20). There was no main effect of convention, (F(1, 590) = 2.19, p = .14).

Figure 2

Combined Severity and Level of Concern Ratings for Admissions Scenario by Error Maker and Convention



Note. When the admissions algorithms were the convention (red bars), algorithm mistakes were judged more severely than human errors, but less so than when human admissions officers were the convention (blue bars). Error bars represent standard error throughout this paper.

However, there was a significant interaction between convention and error maker, F(1, 590) = 6.71, p = .01, η^2_P = 0.01. This interaction indicated that the effect of the identity of the error maker (human or algorithm) on participants' judgments differed depending on whether the error maker was presented as the conventional or non-conventional option. To explore the locus of this interaction, we conducted simple main effects analyses. When the human admissions officers were the conventional option, participants judged the admissions algorithm's mistakes significantly (F(1, 292) = 36.08, p < .001) more severely (M = 5.01, SD = 0.99) compared to mistakes made by the human officers (M = 4.23, SD = 1.22). When admissions algorithms were the convention, participants still judged the algorithm's mistakes (F(1, 298) = 6.08, p = .01) more severely (M = 4.91, SD = 1.01), than mistakes made by the human officer (M = 4.60, SD = 1.16), but this difference was smaller than when the human decision maker was the convention, as indicated by the significant interaction. In short, when the human decision maker was the convention (as in past studies of algorithm aversion), an error made by the algorithm was seen

as much more severe than an identical error made by a human (difference between the blue bars in the figure); but this bias against the algorithm was much less pronounced when the algorithm was the convention (difference between the red bars).

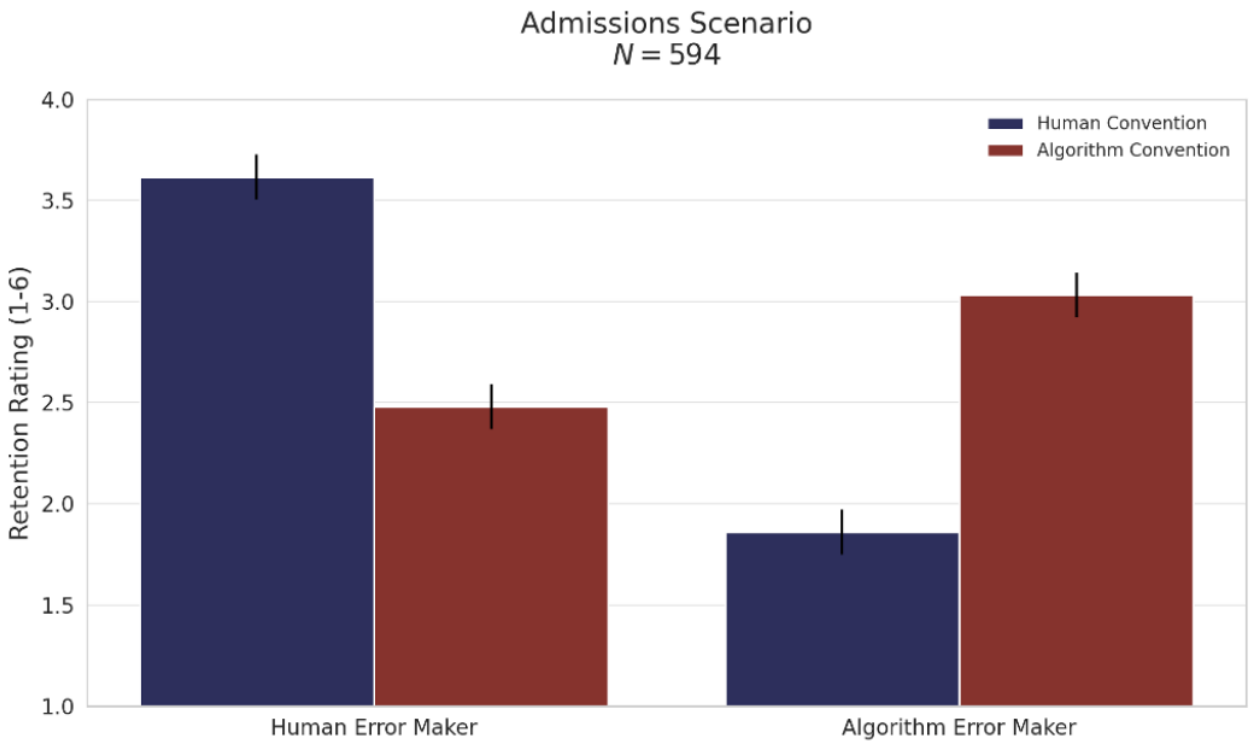
2.2.2. Retention Measure.

The retention question (Table 1) assessed participants' preference for keeping the conventional option used by College X. For instance, it asked whether the college's human admissions officer, Jordan Taylor (in human convention conditions), should be retained, both in the condition where the human was tested and found to make errors and in the condition where the algorithm was tested and found to make errors. As would be expected, participants were much less favorable to retaining the human [algorithm] when the human [algorithm] was tested and found to have made errors (compared to when the human [algorithm] was not tested and the algorithm [human] was tested instead). The main interest here, however, as captured in the analysis above of the severity and concern ratings, was of judgments of the error maker (and how those judgments are influenced by whether the error maker is a human or algorithm, and whether or not the error maker is the convention). For consistency in analysis, we recoded the retention measure so that it consistently reflected a judgment directed toward the error-maker. To do this, the retention measure was reverse-coded for the conditions in which the error was made by the non-conventional option. Following this recoding, in all conditions higher scores indicate a stronger tendency to retain the error maker, despite it having been observed to make errors.

We examined participants' responses using a convention (human convention, algorithm convention) by error maker (human admissions officer, computerized admissions algorithm) ANOVA (Figure 3). We observed a main effect of the error maker, ($F(1, 590) = 29.64$, $p < .001$, $\eta^2_P = 0.05$), indicating that overall, participants more strongly preferred retaining the human admissions officer ($M = 3.04$, $SD = 1.39$) when they made a mistake compared to retaining the admissions algorithm ($M = 2.45$, $SD = 1.54$) when it made identical errors. This suggested that, overall, participants were more forgiving of errors made by the human admissions officer compared to those made by the algorithm. There was no main effect of convention, ($F(1, 590) = 0.03$, $p = .86$).

Figure 3

Preference to Retain Admissions Review Method after it Made an Error



Note. When admissions algorithms were the convention (red bars), participants preferred retaining the algorithm more than the human after identical mistakes. When human admissions officers were the convention (blue bars), participants preferred retaining the human officer over the algorithm.

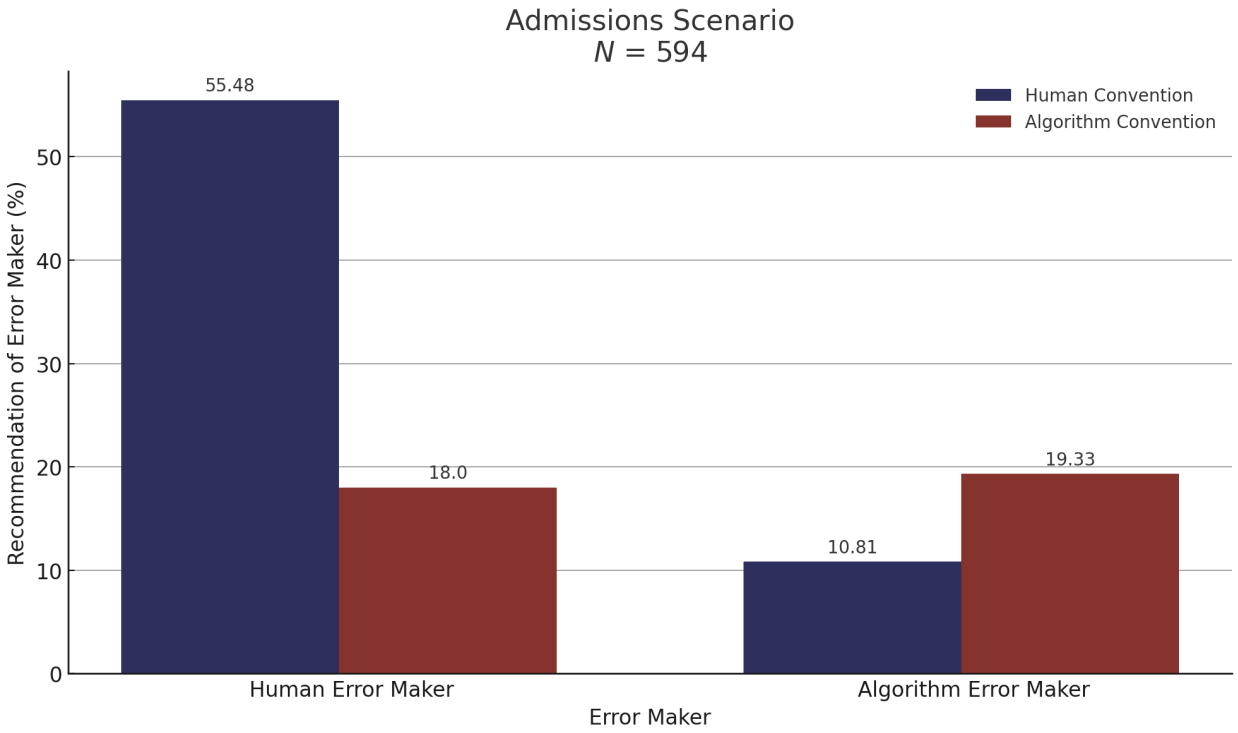
Figure 3 shows that a significant interaction was found between convention and error maker, ($F(1, 590) = 109.09$, $p < .001$, $\eta^2_P = 0.16$), suggesting that the effect of the error maker on participants' retention judgments differed depending on whether the error maker was presented as the conventional or non-conventional option. To unpack this interaction, simple main effects analyses were conducted. We found that participants preferred retaining the human admissions officer more ($M = 3.62$, $SD = 1.30$) than they did the admissions algorithm ($M = 1.86$, $SD = 1.37$) after both made identical mistakes in the human convention condition ($F(1, 292) = 127.96$, $p < .001$; difference between the blue bars in the figure). Conversely, when admissions algorithms were the convention, participants retained the admissions algorithm significantly more ($M = 3.03$, $SD = 1.48$) than the human admissions officer after observing them make the same mistakes ($M = 2.48$, $SD = 1.24$), ($F(1, 298) = 12.35$, $p < .001$; difference between the red bars in the figure).

2.2.3. Recommendation for the Future.

For the final question (Table 1), participants were given a binary choice between recommending human admissions officers or computerized admissions algorithms for use in the future of the state's college admissions. Again, our main interest here—in line with the earlier analyses—was in recommendations for the error maker and how those recommendations were influenced by whether the error maker was a human or an algorithm, and whether or not the error maker was the convention. Hence, to maintain consistency in our analysis, we recoded the recommendation measure to focus on the error maker. This involved creating a new variable that reflected whether the participant had recommended the error maker (coded as 1) or did not recommend the error maker for the future (coded as 0). Following this recoding, Figure 4 shows the proportion of participants who, despite observing mistakes, still recommended the error making agent for future use.

Figure 4

Proportion of Recommendations for Error Makers under Different Conventions in the Admissions Scenario



Note. When admissions algorithms were the convention (red bars), participants recommended both the humans and the algorithms at similar rates after identical mistakes. But when human admissions officers were the convention (blue bars), participants recommended humans more than algorithms.

A Chi-square test was conducted to examine the relationship between the type of error maker (human or algorithm) and participants' recommendation of the error maker (recommended vs. not recommended), specifically within the conditions where the human is the convention (difference between the blue bars in the figure). We found a significant association between the type of error maker and recommendations for error maker in these conditions, (X^2 (1, N = 294) = 66.33, p < .001). This showed that when the human admissions officer was the conventional method for evaluating college applications, participants were significantly more likely to recommend a human error maker (55.48%) compared to an algorithmic error maker (10.81%).

A second Chi-square test was run between the type of error maker and participants' recommendation of the error maker within conditions where the algorithm was the convention (difference between the red bars in the figure). Here, we did not find a significant association between the type of error maker and recommendations for error maker, (X^2 (1, N = 300) = 0.09, p = .767). This result indicated that when the admissions algorithm was the conventional way to screen college applications, participants' likelihood of recommending the error maker did not significantly differ whether the error maker was a human (18.0%) or an algorithm (19.33%).

2.2.4. Summary of Findings

In summary, our results provided evidence for algorithm aversion where expected—when humans were the convention in a human-algorithm comparison. When human admissions officers were the conventional option, participants: judged algorithmic errors more severely; preferred to retain the erring human officer over the erring algorithm; and recommended erring human officers more than the erring algorithms for future use in the state's college admissions. This aligned with our expectations and existing literature on algorithm aversion.

When the admissions algorithm was the conventional choice, however, this pattern no longer held. Specifically, while algorithmic errors were still judged more severely than human errors, the difference was considerably reduced ($M_{\text{difference}}$ = 0.31, SE = 0.13) compared to when humans were the conventional option ($M_{\text{difference}}$ = 0.78, SE = 0.13). Moreover, as seen in Figure 2, human errors were now judged significantly more harshly than they were when human officers were the convention. The main finding here is that when the algorithm was framed as the convention, algorithm aversion was markedly reduced.

For the retention variable, the hypothesis that conventional decision-makers would have higher retention for future tasks than non-conventional ones was strongly supported. Participants showed a higher preference for retaining the admissions algorithm over the human officer when the algorithm was the conventional option—indicating a reversal of the algorithm aversion effect in the other direction. This pattern provided support for our alternate aversion account, suggesting that aversion may be driven not only by the algorithm itself but also by its status as the non-conventional option. Specifically, the drop in retention ratings for the human

error maker when it was the conventional option ($M_{difference}$ = -1.14, SE = 0.16) was similar to the drop for the algorithmic error maker when it was not the conventional option ($M_{difference}$ = -1.18, SE = 0.16).

Finally, for the recommendation for future use, we observed that when the algorithm was the convention—human officers were no longer recommended more often than the algorithm (18.0% vs 19.33%). This can be seen as an elimination of algorithm aversion in future choices when the algorithm is positioned as the conventional option. Furthermore, as seen in Figure 4, there was a significant drop in the recommendation of *human officers* from when they were the convention (55.48% of the time) to when the admissions algorithm was the convention (18.0%).

Changing the context to make the algorithm the conventional option altered participants' judgments and preferences from what we commonly observe in conventional human-algorithm comparisons. We saw algorithm aversion was significantly reduced in judgments of error and reversed or eliminated when participants were asked to make future choices. Overall, these findings provide strong support for our hypothesis that conventionality influences judgments of identical mistakes between humans and algorithms.

3. Study 2

In Study 2, we developed a different hypothetical scenario to test the generalizability of our earlier results. The new scenario involved the quality control of sound speakers, where human and algorithmic agents failed to detect faulty products. In contrast to the forecasting errors in the first study, mistakes made in the present study are detection-based errors. That is, mistakes were shown to be made when either humans or algorithms failed to accurately detect issues within existing products as opposed to forecasting future performance. Since future uncertainty does not play a role in detection-based situations, detection errors might be judged more severely compared to the prediction errors in Study 1. This could be because detection errors—similar to rule-based mistakes where errors result from misapplying existing rules (Reason, 1990)—are seen as more predictable and avoidable than forecasting errors. As in Study 1, the same three characteristics — history, prevalence and system dependence — were used to establish the conventional and alternate options.

3.1. Method

Data and materials for this experiment is available on OSF here: https://osf.io/29sj7/.

3.1.1. Participants

Participants for this study (N = 605, 308 women; 279 men; 6 non-binary; 1 other; $Mdn_{\text{age group}}$ = 35-44) were recruited simultaneously with those for Study 1 through CloudResearch's MTurk Toolkit, using identical recruitment criteria and screening. Participants could only take part in one of the two studies.

3.1.2. Procedure

This experiment was conducted similarly to Study 1, using a fictitious case study focused on the quality control of sound speakers. Participants were briefed about a fictional company specializing in manufacturing sound speakers, which was evaluating two quality control options: a human sound quality analyst and a sound quality computer algorithm. As in Study 1, either the human or the algorithmic agent was presented as the conventional method for quality control (detailed in Figure 5). During testing by the company, either the human analyst or the algorithm was shown to have failed in detecting 40% of the faulty speakers. The remainder of the experimental procedure mirrored Study 1, including comprehension checks, judgment questions (Table 1), and demographics questions.

Figure 5

Sound Speakers Scenario as Observed by Participants

Part 1
<u>Case Background</u> Company X is one of several companies in the country that manufacture large sound speakers used in private and commercial settings. The company is currently solidifying its strategic plan for the next five years. Part of this involves reviewing its quality control methods. Specifically, it's assessing two different options for testing its speakers: a *human* sound quality analyst or a *computerized algorithm* designed to analyze sound quality.
Part 2
Case Details For the past ten years, the company has [employed Jordan Taylor, a human analyst, to test the sound quality of its speakers / used SonicVerifier 3.0, a computer algorithm, to test the sound quality of its speakers.] These days, most manufacturers (around *85%*) use [an analyst, while the rest use a computer program to test their speakers / a computer program, while the rest use a human analyst to test their speakers.] As it can take a lot of time and resources to [develop custom-built sound-quality computer algorithms, companies across the country currently depend on human sound-quality analysts, like Jordan Taylor, to keep their operations running / train specialist human sound-quality analysts, companies across the country currently depend on sound-quality computer algorithms, like SonicVerifier, to keep their operations running.]
Part 3
The company decides to run multiple tests on both options before making a final decision on its quality control method.
Part 4
Test: Human Sound Quality Analyst In one test, the company tasked Jordan Taylor, [their sound-quality analyst /a human sound-quality analyst employed in the industry], to assess 100 previously manufactured

speakers. Out of these, 30 speakers had failed prior quality checks. These 30 speakers were deliberately inserted into the 100 assessed, for testing purposes.

Taylor was then asked to **find the 30 speakers from the 100** provided that should fail a quality control test.

When the test results were reviewed, the company found that Taylor had **missed 12 of the 30 malfunctioning speakers** that had failed previous quality control tests.

OR

Test: Sound Quality Computer Algorithm

In one test, the company tasked SonicVerifier 3.0, [a sound-quality computer algorithm used in the industry / their sound-quality algorithm], to assess 100 previously manufactured speakers. Out of these, 30 speakers had failed prior quality checks. These 30 speakers were deliberately inserted into the 100 assessed, for testing purposes.

SonicVerifier was then asked to **find the 30 speakers from the 100** provided that should fail a quality control test.

When the test results were reviewed, the company found that SonicVerifier had ***missed 12 of the 30 malfunctioning speakers*** *that had failed previous quality control tests.*

Note. This includes all scenario parts of Study 2 in sequence, as seen by the participants, barring the comprehension questions. Colored text within square brackets shows variations made per condition (blue = human convention, red = algorithm convention).

3.2. Results

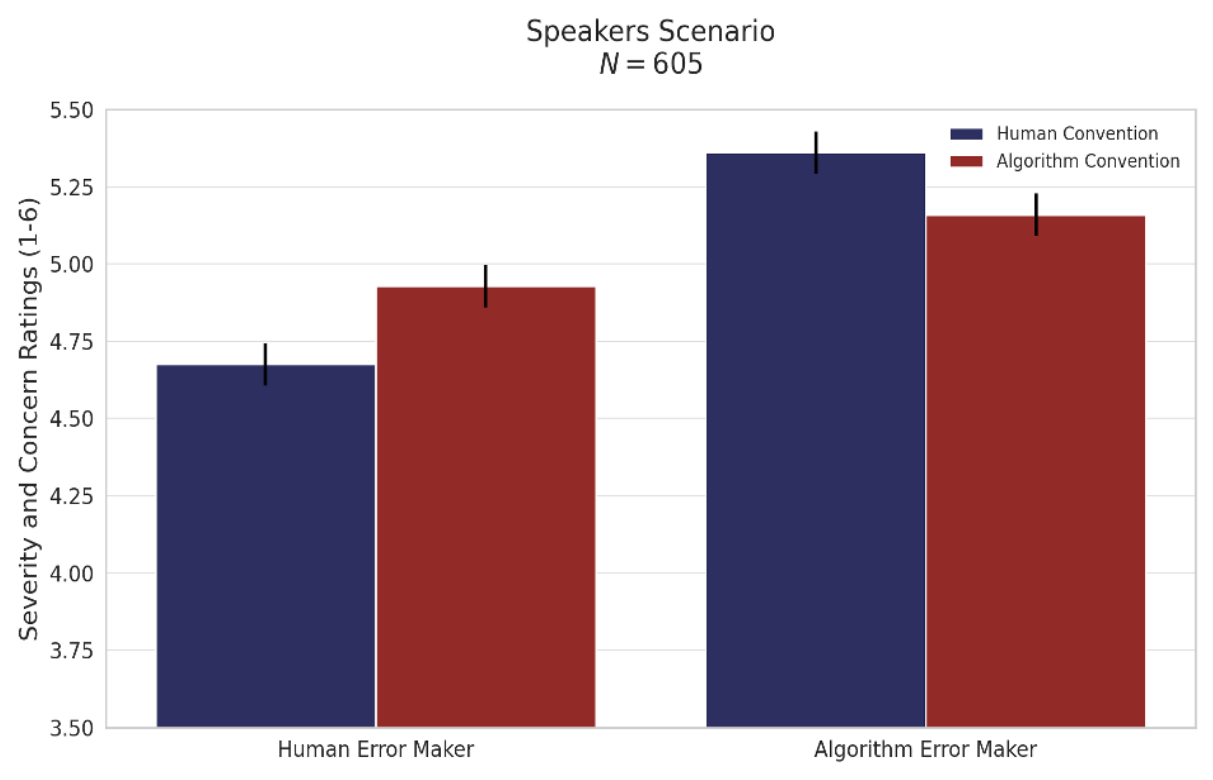
3.2.1. Mistake Severity & Concern Measure.

Recall that in Study 1, we conducted a correlation analysis to examine the relationship between mistake severity and level of concern ratings. Due to a high correlation, we created a new combined variable by averaging the two variables across all participants. Similarly, in Study 2, there was a strong correlation between mistake severity and concern measures, $r(603) = .735$, $p < .001$, 95% CI [0.696, 0.770], prompting the creation of an identical combined variable for this analysis.

We examined participants' responses using a convention (human convention, algorithm convention) by error maker (human analyst, sound quality algorithm) ANOVA (Figure 6). There was a main effect of the error maker, ($F(1, 601) = 44.21$, $p < .001$, $\eta^2_P = 0.07$), indicating that overall, participants judged the mistakes of the computerized quality control algorithm more severely ($M = 5.26$, $SD = 0.73$) than they did the mistakes of the human analyst ($M = 4.80$, $SD = 0.97$). Again, there was no main effect of convention, ($F(1, 601) = 0.14$, $p = .71$).

Figure 6

Combined Severity and Level of Concern Ratings for Speakers Scenario by Error Maker and Convention



Note. As in Study 1, when the sound quality algorithms were the convention (red bars), algorithm mistakes were judged more severely than human errors, but less so than when human analysts were the convention (blue bars).

As in Study 1, there was a significant interaction between convention and error maker, (F(1, 601) = 10.81, p = .001, η^2_P = 0.02). This interaction showed that the effect of the identity of the error maker (human or algorithm) on participants' judgments varied depending on whether the error maker was presented as the conventional or non-conventional option. To explore the locus of this interaction, we conducted simple main effects analyses. When the human sound quality analysts were the conventional option, participants judged the sound quality algorithm's mistakes significantly (F(1, 300) = 50.39, p < .001) more severely (M = 5.36, SD = 0.65) compared to mistakes made by the human officers (M = 4.68, SD = 0.99). When sound quality algorithms were the convention, participants still judged the algorithm's mistakes (F(1, 301) = 5.54, p = .012) more severely (M = 5.16, SD = 0.78), than mistakes made by the human officer (M = 4.93, SD = 0.93). Similar to Study 1, this difference was smaller than when the human analyst was the convention, as indicated by the significant interaction noted above. Overall, we saw again that when the human decision maker was the convention (as in past work on

algorithm aversion), an error made by the algorithm was seen as much more severe than an identical error made by a human (difference between the blue bars in Figure 6); but this bias against the algorithm was much less pronounced when the algorithm was the convention (difference between the red bars).

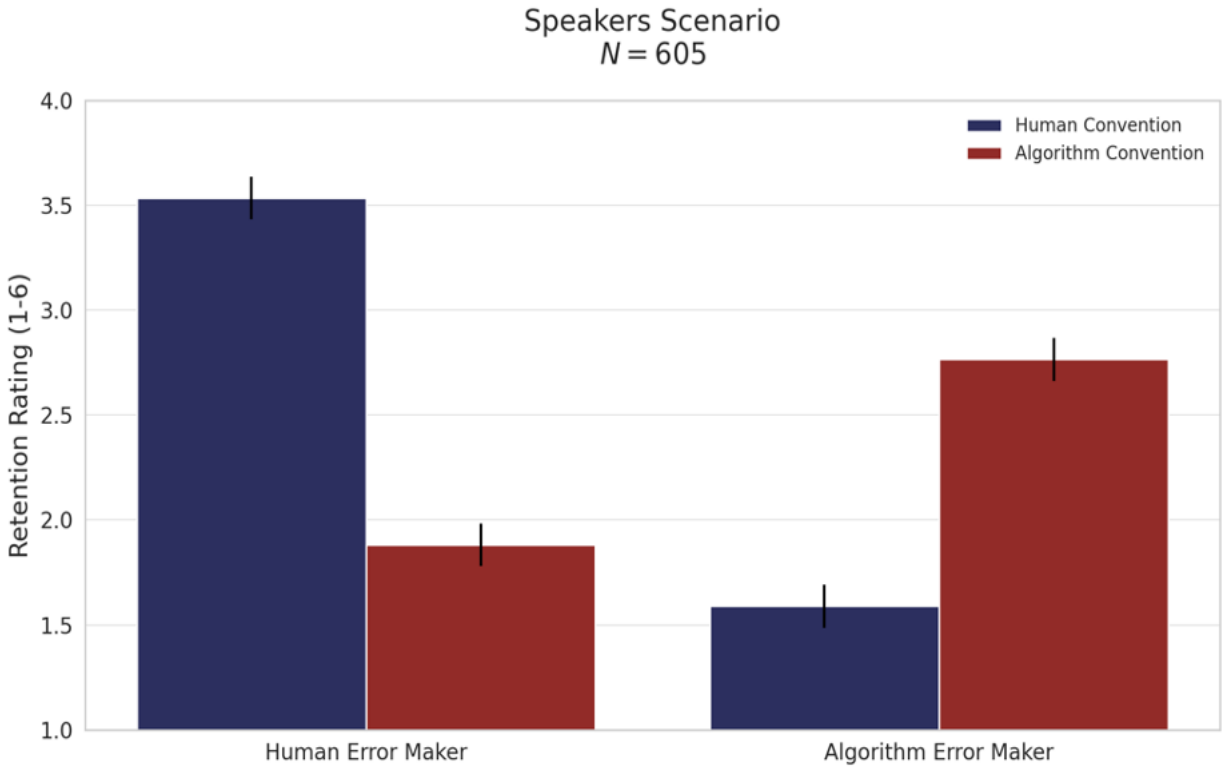
3.2.2. Retention Measure.

The retention question (Table 1) assessed participants' preference for keeping the conventional option used by Company X. For instance, it asked whether the college's human sound quality analyst, Jordan Taylor (in human convention conditions), should be retained, both in the condition where the human was tested and found to make errors and in the condition where the algorithm was tested and found to make errors. As was expected for the retention measure in Study 1, participants were much less favorable to retaining the human [algorithm] when the human [algorithm] was tested and found to have made errors (compared to when the human [algorithm] was not tested and the algorithm [human] was tested instead). The main interest here remained the judgments of the error maker (and how those judgments are influenced by whether the error maker is a human or algorithm, and whether or not the error maker is the convention). To maintain consistency in the analysis, we adjusted the retention measure to uniformly reflect judgments directed at the error-maker. This involved reverse-coding the retention measure for conditions where the error was made by the non-conventional option. After this recoding, higher scores across all conditions indicated a stronger tendency to retain the error maker, despite it having been observed to make mistakes.

We examined participants' responses using a convention (human convention, algorithm convention) by error maker (human analyst, sound quality algorithm) ANOVA (Figure 7). We observed a main effect of the error maker, ($F(1, 601) = 42.70$, $p < .001$, $\eta^2_P = 0.04$), indicating that overall, participants preferred retaining the human analyst ($M = 2.70$, $SD = 1.48$) more when they made a mistake than they preferred retaining the sound quality algorithm ($M = 2.18$, $SD = 1.41$) when it made identical errors. This implied that participants were, overall, more forgiving of errors made by the human analyst compared to the algorithm. This time, we observed an unexpected main effect of convention, ($F(1, 601) = 5.45$, $p = .02$, $\eta^2_P = 0.01$), which indicated that participants had a slightly higher preference for retaining error makers in the human convention conditions ($M = 2.56$, $SD = 1.57$) than they did in the algorithm convention conditions ($M = 2.32$, $SD = 1.35$). We do not consider this relevant for our current analyses.

Figure 7

Preference to Retain Sound Testing Method after it Made an Error



Note. Similar to Study 1, when sound quality algorithms were the convention (red bars), participants retained the algorithm more than the human after identical mistakes. When human analysts were the convention (blue bars), participants preferred retaining the human over the algorithm.

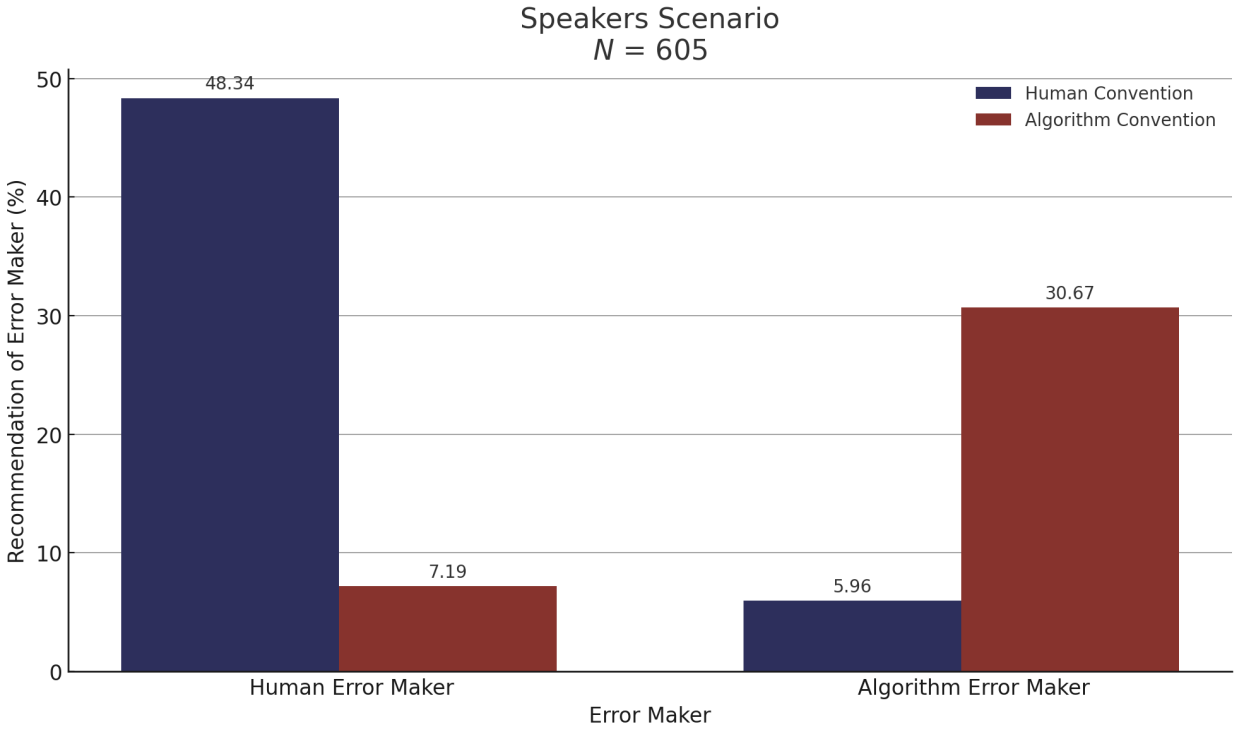
A significant interaction between convention and error maker was found, ($F(1, 601) = 192.0$, $p < .001$, $\eta^2_P = 0.24$), indicating that the impact of the error maker on participants' retention judgments varied considerably based on whether the error maker was the conventional or non-conventional option. To further explore this interaction, simple main effects analyses were conducted. In the human convention condition, participants preferred retaining the human analyst($M = 3.54$, $SD = 1.29$) significantly more than the admissions algorithm ($M = 1.59$, $SD = 1.17$) after both made identical mistakes, ($F(1, 300) = 187.44$, $p < .001$; difference between the blue bars in Figure 7). Conversely, in the algorithm convention condition, ($F(1, 301) = 36.33$, $p < .001$; difference between the red bars in the figure), participants were more likely to retain the sound quality algorithm ($M = 2.77$, $SD = 1.39$) than the human analyst ($M = 1.88$, $SD = 1.16$) after observing the same mistakes.

3.2.3. Recommendation for the Future.

For the final question of Study 2 (Table 1), participants were asked to recommend whether Company X should use a human sound-quality analyst or a computerized sound-quality algorithm for the next five years. As in Study 1, our main interest here—in line with the earlier analyses—was in recommendations for the error maker and how those recommendations were influenced by whether the error maker was a human or an algorithm, and whether or not the error maker was the convention. To maintain consistency in our analysis, we recoded the recommendation measure to focus on the error maker. As earlier, this involved creating a new variable that reflected whether the participant had recommended the error maker (coded as 1) or did not recommend the error maker for the future (coded as 0). Following this recoding, Figure 8 shows the proportion of participants who, despite observing mistakes, still recommended the error making agent for future use.

Figure 8

Proportion of Recommendations for Error Makers under Different Conventions in the Speakers Scenario



Note. When sound quality algorithms were the convention (red bars), participants recommended the algorithm more than the human for future use after identical mistakes. When human analysts were the convention (blue bars), participants recommended humans more than algorithms.

As in Study 1, a Chi-square test was conducted to examine the relationship between the type of error maker (human or algorithm) and participants' recommendation of the error maker (recommended vs. not recommended), specifically within the conditions where the human is the convention (difference between the blue bars in Figure 8). There was a significant association between the type of error maker and recommendations for error maker in these conditions, (X^2 (1, N = 302) = 68.57, p < .001). Consistent with Study 1 results, this showed that when the human analyst was the conventional method for testing sound speakers, participants were significantly more likely to recommend a human error maker (48.34%) compared to an algorithmic error maker (5.96%).

A second Chi-square test was run between the type of error maker and participants' recommendation of the error maker within conditions where the algorithm was the convention (difference between the red bars in the figure). This time, we found a significant association between the type of error maker and recommendations for error maker under these conditions, (X^2 (1, N = 303) = 27.33, p < .001). As in Study 1, we saw the aversion to algorithms was eliminated when the algorithm was the convention, with human analysts no longer recommended more often than the algorithms. In fact, in this experiment, the effect reversed, such that the algorithms were favored over the human analysts (30.67% vs 7.19%) for future recommendations.

3.2.4. Summary of Findings

In line with Study 1, results from Study 2 provided evidence for algorithm aversion where we expected it—when humans were the convention in a human-algorithm comparison. When human sound quality analysts were the conventional option, participants: judged algorithmic errors more severely; preferred to retain the erring human analyst over the erring algorithm; and recommended erring human analysts more than the erring algorithms for future use by the company. Again, this aligned with our expectations and existing literature on algorithm aversion.

The overall patterns in the severity and concern ratings in Study 2 were very similar to Study 1 results. In Study 2, we again observed that while algorithmic errors were still judged more severely than human errors, the difference was considerably reduced ($M_{\text{difference}}$ = 0.23, SE = 0.10) compared to when humans were the conventional option ($M_{\text{difference}}$ = 0.69, SE = 0.10). We also see in Figure 6 that human errors were judged significantly more harshly when the sound quality algorithm was the convention, than when the human analysts were the convention. The main finding of interest here is that when the sound quality algorithm was framed as the convention, algorithm aversion was markedly reduced.

For the retention variable, similar to Study 1, our hypothesis that conventional decision-makers would have higher retention for future tasks than non-conventional ones was strongly supported. Participants showed a higher preference for retaining the sound quality algorithm over the human analyst when the algorithm was the conventional option—showing a reversal of the algorithm aversion effect in the other direction. As in Study 1, this provided support for our alternate aversion account. In fact, in Study 2, the drop in retention ratings for the human error maker when the human analyst was the conventional option to when it was not ($M_{difference}$ = -1.66, SE = 0.14) was larger, compared to the drop in retention for the algorithmic error maker from when the sound quality algorithm was conventional option to when it was not ($M_{difference}$ = -1.18, SE = 0.15). This large drop in retention for the human analyst showed that when humans were the alternate option, tolerance for their mistakes fell considerably.

Results for the recommendations for future use variable showed that, like Study 1, when the algorithm was the convention, human analysts were no longer recommended more often than the algorithms. In Study 2, this went even further and algorithm aversion reversed as we saw algorithms preferred over the human analysts (30.67% vs 7.19%) for future recommendations after making identical mistakes. This provides further support for our alternate aversion account.

Overall, the results of Study 2 indicate that our findings from Study 1 are generalizable to a different contextual situation. We found strong support for our hypothesis that participants' evaluations of mistakes are influenced by the conventional status of the error maker. Specifically, changing the context to make the algorithm the conventional option significantly altered participants' judgments and preferences, reducing, eliminating or even reversing algorithm aversion that is commonly observed in traditional human-algorithm comparisons.

4. General Discussion

It has long been known that the status quo really matters (Samuelson & Zeckhauser, 1988): default choices can shape subsequent judgments and behaviors (Johnson & Goldstein, 2003). This work provides strong evidence that biases stemming from the status quo also go on to have a consequential influence on people's evaluations of identical mistakes. Specifically, in this paper, we operationalized the status quo as conventional options and investigated how the conventional status of the error maker influences judgments and decisions. In both studies, we observed that when human admissions officers or human analysts were considered the conventional choice, identical errors made by algorithms were judged more harshly compared to those made by a human. Put simply, participants generally judged algorithmic mistakes more severely than human mistakes. This aligns with previous research on algorithm aversion, where people exhibit a distrust in algorithmic decision-making.

However, it is also relevant to consider that people have an inherent bias towards maintaining the conventional status quo, which traditionally frames humans as the conventional choice in human-algorithmic comparisons. Perhaps this bias can lead to a harsher evaluation of errors made by non-conventional algorithmic decision-makers compared to their human

counterparts. To test this idea, we changed the equation and found that when we framed the algorithm as the conventional option, there was a significant impact on people's judgments and choices. This change diminished, eliminated or even reversed the usual algorithm aversion observed in standard human-algorithm comparisons.

Alternate Aversion. The results support our proposed idea of alternate aversion, where we hypothesized that the presence of a conventional option will lead to a stronger aversion for the non-conventional option. They suggest that harsher judgments or treatments of the alternate options might be in response to a perceived threat to the status quo. In the case of human-algorithm comparisons, where the human has often been situated as the conventional choice, perhaps it is not just *algorithm* aversion that is observed—but also an aversion to the algorithm due to its alternate status, contributing to its harsher treatment. To support this idea, it was important to show that putting another agent, a human in this case, in the alternate position would yield similar results. We found that participants' judgments were indeed significantly influenced by the conventional status of the error maker. When the algorithm was positioned as the conventional method and the humans as the alternate—things changed. Human mistakes were judged less leniently, the disparity in severity judgments between algorithmic and human mistakes decreased significantly, and in one study people made future choices that aligned with keeping the status quo, which in this case was algorithms. This also indicates that conventionality mitigated some bias against algorithms. Importantly, it suggests that the context of conventionality, rather than algorithm aversion alone, contributes to a bias that is often primarily attributed to a variety of algorithm-centric traits.

The reasons for algorithm aversion seem less potent and explanative today then they did 30 years ago. One only needs to take a brief glimpse at the world around to see that, despite the litany of concerns about the use of algorithms and AI, their outsized presence continues to expand into most facets of human life. We, like others, suspect that "the hostility to algorithms will probably soften as their role in everyday life continues to expand" (Kahneman, 2011). We see this coming about in different ways. In research, an inverse phenomenon called *algorithm appreciation* has emerged, where people show a higher preference for algorithms over humans in certain contexts. In a series of experiments, Logg et al. (2019) found that participants consistently preferred algorithmic advice over human advice, using it to improve their judgment accuracy in tasks like numeric estimates, song popularity forecasts, and romantic attraction predictions. In a separate study involving 157 medical doctors (GPs) using a validated cancer risk calculator, doctors weighed their own judgments and the algorithm equally. Algorithmic advice remained influential even when it conflicted with their' judgments. These findings challenge the notion that people would tend to favor their own judgment, suggesting that algorithmic advice can substantially influence clinical judgments (Pálfi et al., 2022). In another study involving stock market predictions, researchers found that familiarity with algorithms increased trust, which in turn drove algorithm appreciation. Additionally, a high-performing algorithm strengthened the effect of familiarity with algorithms on trust (Mahmud et al., 2024). Even on issues such as discrimination in hiring practices, Bigman and colleagues (2023) have reported that people are actually less morally outraged by algorithmic discrimination than by human discrimination.

Overall, there seems to be an ongoing shift in attitudes towards algorithms. How do findings like these align with the literature on algorithm aversion? One explanation suggests that algorithm aversion and appreciation can coexist, that people initially show algorithm appreciation, preferring algorithmic advice over human advice—before they receive feedback on the algorithm's performance. Once they learn the algorithm has made an error, algorithm aversion sets in, and they are more likely to prefer human judgment even when it is inferior (Dietvorst et al., 2015; Logg, 2022). Our research suggests that under certain contexts, people can actually continue preferring the algorithm despite seeing it make an error, and this is most likely to occur when the algorithm is the conventional decision-maker. This is not to say that our results contradict the idea that algorithm appreciation can be followed by algorithm aversion when an error is made. Rather, our results show that the convention of the day matters, and human reactions to algorithmic mistakes may not be as predictable as previously believed when algorithms are the conventional method in that context.

Judgments vs. Choices. In our studies, there were two types of judgment—one about how severe the error was; the other about which agent should be retained and recommended for the future, in light of the error. Across both studies, when the human is the convention, the usual algorithm aversion appears in both judgments of mistake severity and recommendations of which system to use going forward. When the algorithm is the convention, the algorithm's mistakes are not judged as less severe than the human's, but now recommendations about which to retain do tend to favor the algorithm. In short, the effect of the convention appears to be stronger on these future recommendations than on judgments of mistake severity, where algorithm aversion seemed to be more persistent.

One way to interpret this is that when the algorithm is the convention, people still judge algorithmic errors quite severely, but this bias was overcome or mitigated when it came to actionable decision-making. This finding seemingly contradicts research that has shown that once people see an algorithm make an error, they trust it less than a human agent in the future, even when the algorithm has better accuracy overall (Dietvorst et al., 2015). We do not think that is a contradiction to earlier research. Instead, it reiterates the fact that in most human-algorithm comparisons, it is rarely considered whether people view the algorithm as the norm for that task, nor is the algorithm explicitly established as the conventional choice, as we did in these experiments. An argument can be made that this preference for the convention is rational, regardless of whether it leads to a bias for human or algorithmic error makers. We described convention based on three characteristics: familiarity and history with the agent, prevalence of the agent's use in the surrounding environment, and dependence of the system on the agent. Based on these factors, if an agent is the conventional method of doing things and it errs, penalizing it severely or discontinuing its use can be costly. Therefore, it is rational to not discard it right away. While people might feel a bias against the algorithm for all the reasons that have been historically proposed, they are likely to be more objective in their actual decisions. It is also noteworthy that while earlier research showed that superior accuracy was not enough for people to prefer algorithms over humans after mistakes, this work has shown that the algorithmic agent being the convention could be a more persuasive factor than objective performance. This is a potential area for future research, where the performance and

conventionality of error makers could be further explored, particularly in how these factors interact to influence trust and decision-making.

This duality in judgments and actionable decisions also suggests that people's aversion to algorithms after they make mistakes should not be considered an outright predictor of their actual decisions. If the algorithm is the conventional option for a task, people may choose to persist with it even after it has made mistakes. We can expect to see more of these trends as people's lives become increasingly intertwined with algorithms, and they become the conventional method for more tasks. Through our digital worlds, we are now developing the history and familiarity with algorithms that were not present in the 20th century, and the use of similar and shared algorithmic applications is widely prevalent amongst our social groups and work environments. Furthermore, for certain algorithmic applications such as Google Maps, one can even argue that the system—everyday navigation in this case—is highly dependent on them, and making mistakes would not cause people to abandon them right away. The extent of the persistence to continue with error makers can vary depending on the decision context.

Prediction vs. detection-based differences. The distinction between prediction-based and detection-based errors might also have influenced participants' judgments to an extent. In Study 1, errors were related to predicting future outcomes (admissions decisions), which inherently involve uncertainty and complexity. People might be more lenient towards prediction-based errors due to the elements of inherent uncertainty in forecasting the future, especially in the case of human predictions (Ganzach & Krantz, 1991). For instance, participants could attribute such errors to better-than-expected performance by applicants during their college years resulting from unaccounted-for factors like personal growth or changes in motivation due to new experiences (Deci & Ryan, 2000; Kilgo et al., 2016). In contrast, detection errors might be viewed as more avoidable and thus judged more harshly. This context could have potentially contributed to why human analysts in the sound speaker scenario (Study 2)were judged more harshly relative to algorithms, compared to human admissions officers in the first study.

Human outcomes. In Study 1, participants observed the admissions officer or the algorithm make an error in forecasting the future, which had the potential to directly and negatively affect human applicants when deployed for admissions screening. People might be more opposed to algorithmic or AI-based judgments when these decisions directly impact human lives, as opposed to situations with non-human outcomes (Grove & Meehl, 1996; Schlund & Zitek, 2021). In Study 2, we used a different hypothetical scenario that did not have direct human outcomes—instead, an inanimate object (a sound speaker) was at the wrong end of the faulty decision-making. Some differences in results between the two studies, such as lower future recommendation rates for the algorithmic error maker in the admissions scenario compared to the sound speakers scenario, could perhaps be partly attributed to this factor. These results support the need for more contextual distinction and research into people's evaluations of and reactions to mistakes, depending on who or what is affected by the errors.

Research contributions. Research on algorithm aversion originates from a variety of different perspectives and disciplines. Each field approaches algorithm aversion with a unique—

though often overlapping—set of questions. For instance, an economics perspective might explore: *To what extent does algorithm aversion affect consumer behavior in markets where algorithmic recommendations are common?* From a developer and design perspective, the focus could be on: *What design features of algorithmic systems—such as interface design and explainability—influence user acceptance or rejection of these technologies?* In organizational behavior, key questions might revolve around: *What is the impact of algorithm aversion on employee hiring, performance evaluations, job satisfaction, and the adoption of algorithmic tools in the workplace?* Legal studies might ask: *How does algorithm aversion heighten ethical concerns in areas like criminal justice, policing, and sentencing—where perceptions of fairness and transparency are paramount?* The healthcare sector might be interested in*: What barriers does algorithm aversion create for the adoption of highly accurate algorithmic technologies in diagnostic settings?* While psychological research might ask: *What cognitive biases influence people's trust or distrust in algorithms?*

The end goals of these questions can vary greatly—some might be motivated to maximize profits, others seek to overcome resistance to tools that have clear potential to improve human well-being (Longoni et al., 2019), and still others strive to use this modern phenomenon to gain insights into the human psyche. Our research fits into this last category—it seeks to further our understanding of how people judge and respond to mistakes, particularly within the contemporary context of human versus algorithmic decision-making. Yet, as research across various fields has increasingly converged on algorithm aversion, the collective experimental results have become more complex and, at times, seemingly contradictory (Jussupow et al., 2020; Logg, 2022). This raises important questions about when we can reliably predict algorithm aversion or appreciation or explain why they occur in the first place.

In this regard, this research makes two important contributions. First, it provides evidence that algorithm aversion is not a deeply ingrained cognitive bias on the same plane as confirmation bias or the availability heuristic. As our studies showed, algorithm aversion diminishes or even reverses direction when the algorithm becomes the conventional choice. This finding emphasizes that algorithm aversion is highly context-dependent rather than a universal bias. Second, this work suggests that algorithm aversion or appreciation are closely linked to longstanding human cognitive biases. By examining people's interactions with algorithms through the lenses of the status quo bias, choice architecture, and other established decision-making frameworks discussed earlier, the research has shown that the conventionality status of the algorithm can play a substantial role in predicting whether people will exhibit aversion to or appreciation for it. Put simply, algorithm aversion might not be as pervasive or complex as it initially appears, and a human preference for the existing status quo could be a key factor in explaining and predicting human interactions with algorithms.

Limitations and Conclusion. Our research offers important insights into algorithm appreciation and aversion, yet it is not without its limitations. Firstly, the studies relied heavily on controlled hypothetical scenarios which, while useful for isolating variables, may not fully capture the complexity of real-world decision-making environments. Future research could aim to replicate these findings in more naturalistic settings—such as actual interactions with an algorithmic agent (e.g., an AI chatbot) to establish its conventionality—to enhance ecological validity. To explore the effects of conventionality, we focused on specific tasks—college admissions and quality control for sound speakers—and varied the type of task—prediction-based and detection-based. Further studies should explore a broader range and types of tasks to determine the generalizability of our findings.

Our notion of 'conventionality' is rooted in existing work on cognitive biases, we developed the idea of alternate aversion, and we showed that conventionality plays a significant role in mitigating people's reactions to algorithmic errors. However, we still did not cut deeply to the underlying psychological mechanisms. Understanding the psychological mechanisms behind conventionality in such decision-making contexts will provide deeper insights into how and why it influences our preferences and aversions in situations where we have established norms and expectations. Further research can go into which of the three factors explored here—history, prevalence, system dependence—as well as other potential factors, are most relevant in establishing conventionality and inducing aversion for the alternate option.

Finally, this research suggests that algorithm aversion might not be the stable bias it is often thought to be. Our findings indicate that algorithm aversion can be significantly influenced by the context in which the algorithm operates, particularly when it is framed as the conventional option. This context-dependent nature of algorithm aversion implies that with appropriate framing and integration, it is perhaps possible to mitigate or even reverse this bias. This can start with understanding factors such as the perceived status quo of that specific domain. Additionally, the way algorithms are introduced and integrated into existing systems—such as whether they are seen as tools or replacements for human decision-makers—could influence trust and acceptance. Understanding these dynamics will not only contribute to the theoretical framework of human-algorithm interaction but also inform the design and implementation of algorithmic systems in various domains.

References

Alang, N. (2016, May 13). Life in the Age of Algorithms. *New Republic*. https://newrepublic.com/article/133472/life-age-algorithms

Baron, J., & Ritov, I. (2004). Omission bias, individual differences, and normality. *Organizational Behavior and Human Decision Processes*, *94*(2), 74–85. https://doi.org/10.1016/j.obhdp.2004.03.003

Britannica, T. Editors of Encyclopaedia (2024, May 24). al-Khwārizmī. Encyclopedia Britannica. https://www.britannica.com/biography/al-Khwarizmi

Bigman, Y. E., & Gray, K. (2018). People are averse to machines making moral decisions. *Cognition*, *181*, 21–34. https://doi.org/10.1016/j.cognition.2018.08.003

Bigman, Y. E., Wilson, D., Arnestad, M. N., Waytz, A., & Gray, K. (2023). Algorithmic discrimination causes less moral outrage than human discrimination. *Journal of Experimental Psychology: General*, *152*(1), 4–27. https://doi.org/10.1037/xge0001250

Bonezzi, A., Ostinelli, M., & Melzner, J. (2022). The human black-box: The illusion of understanding human better than algorithmic decision-making. *Journal of Experimental Psychology: General*, *151*(9), 2250–2258. https://doi.org/10.1037/xge0001181

Bonezzi, A., & Ostinelli, M. (2021). Can algorithms legitimize discrimination? *Journal of Experimental Psychology: Applied*, *27*(2), 447–459. https://doi.org/10.1037/xap0000294

Brzezicki, M. A., Bridger, N. E., Kobetić, M. D., Ostrowski, M., Grabowski, W., Gill, S. S., & Neumann, S. (2020). Artificial intelligence outperforms human students in conducting

neurosurgical audits. *Clinical Neurology and Neurosurgery*, *192*, 105732. https://doi.org/10.1016/j.clineuro.2020.105732

Carabantes, M. (2020). Black-box artificial intelligence: An epistemological and critical analysis. *AI & SOCIETY*, *35*(2), 309–317. https://doi.org/10.1007/s00146-019-00888-w

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

Chong, L., Zhang, G., Goucher-Lambert, K., Kotovsky, K., & Cagan, J. (2022). Human confidence in artificial intelligence and in themselves: The evolution and impact of confidence on adoption of AI advice. *Computers in Human Behavior*, *127*, 107018. https://doi.org/10.1016/j.chb.2021.107018

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist*, *34*(7), 571–582. https://doi.org/10.1037/0003-066X.34.7.571

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical Versus Actuarial Judgment. *Science*, *243*(4899), 1668–1674. https://doi.org/10.1126/science.2648573

Deci, E. L., & Ryan, R. M. (2000). The "What" and "Why" of Goal Pursuits: Human Needs and the Self-Determination of Behavior. *Psychological Inquiry*, *11*(4), 227–268. https://doi.org/10.1207/S15327965PLI1104_01

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err. *Journal of Experimental Psychology. General*, *144*(1), 114–126. https://doi.org/10.1037/xge0000033

Dinner, I., Johnson, E. J., Goldstein, D. G., & Liu, K. (2011). Partitioning default effects: Why people choose not to choose. *Journal of Experimental Psychology: Applied*, *17*(4), 332–341. https://doi.org/10.1037/a0024354

Einhorn, H. J. (1986). Accepting Error to Make Less Error. *Journal of Personality Assessment*, *50*(3), 387–395. https://doi.org/10.1207/s15327752jpa5003_8

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines*, *28*(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

Eastwood, J., Snook, B., & Luther, K. (2012). What People Want From Their Professionals: Attitudes Toward Decision-making Strategies. *Journal of Behavioral Decision Making*, *25*(5), 458–468. https://doi.org/10.1002/bdm.741

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. https://doi.org/10.1037/1076-8971.2.2.293

Highhouse, S. (2008). Stubborn Reliance on Intuition and Subjectivity in Employee Selection. *Industrial and Organizational Psychology*, *1*(3), 333–342. https://doi.org/10.1111/j.1754-9434.2008.00058.x

Holm, E. A. (2019). In defense of the black box. *Science*, *364*(6435), 26–27. https://doi.org/10.1126/science.aax0162

Johnson, E. J., & Goldstein, D. (2003). Do Defaults Save Lives? *Science*, *302*(5649), 1338–1339. https://doi.org/10.1126/science.1091721

Johnson, E. J., Shu, S. B., Dellaert, B. G. C., Fox, C., Goldstein, D. G., Häubl, G., Larrick, R. P., Payne, J. W., Peters, E., Schkade, D., Wansink, B., & Weber, E. U. (2012). Beyond nudges: Tools of a choice architecture. *Marketing Letters*, *23*(2), 487–504. https://doi.org/10.1007/s11002-012-9186-1

Jussupow, Ekaterina; Benbasat, Izak; and Heinzl, Armin, "Why Are We Averse Towards Algorithms? A Comprehensive Literature Review On Algorithm Aversion" (2020). *In Proceedings of the 28th European Conference on Information Systems (ECIS)*, An Online AIS Conference, June 15-17, 2020. https://aisel.aisnet.org/ecis2020_rp/168

Kahneman, D., Knetsch, J. L., & Thaler, R. H. (1991). Anomalies: The Endowment Effect, Loss Aversion, and Status Quo Bias. *Journal of Economic Perspectives*, *5*(1), 193–206. https://doi.org/10.1257/jep.5.1.193

Kilgo, C. A., Mollet, A. L., & Pascarella, E. T. (2016). The Estimated Effects of College Student Involvement on Psychological Well-Being. *Journal of College Student Development*, *57*(8), 1043–1049. https://doi.org/10.1353/csd.2016.0098

Logg, J. M. (2022). The psychology of Big Data: Developing a "theory of machine" to examine perceptions of algorithms. In *The Psychology of Technology* (pp. 349–378). American Psychological Association. https://doi.org/10.1037/0000290-011

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

Longoni, C., Bonezzi, A., & Morewedge, C. K. (2019). Resistance to Medical Artificial Intelligence. *Journal of Consumer Research*, *46*(4), 629–650. https://doi.org/10.1093/jcr/ucz013

Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and review of the literature*. Minneapolis, MN: University of Minnesota Press.

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. https://doi.org/10.1002/bdm.637

Pálfi, B., Arora, K., & Kostopoulou, O. (2022). Algorithm-based advice taking and clinical judgement: Impact of advice distance and algorithm information. *Cognitive Research: Principles and Implications*, *7*(1), 70. https://doi.org/10.1186/s41235-022-00421-6

Promberger, M., & Baron, J. (2006). Do patients trust computers? *Journal of Behavioral Decision Making*, *19*(5), 455–468. https://doi.org/10.1002/bdm.542

Rainie, L., & Anderson, J. (2017, February 8). *Code-Dependent: Pros and Cons of the Algorithm Age*. Pew Research Center: Internet, Science & Tech. https://www.pewresearch.org/internet/2017/02/08/code-dependent-pros-and-cons-of-the-algorithm-age/

Ritov, I., & Baron, J. (1992). Status-quo and omission biases. *Journal of Risk and Uncertainty*, *5*(1). https://doi.org/10.1007/BF00208786

Samuelson, W., & Zeckhauser, R. (1988). Status quo bias in decision making. *Journal of Risk and Uncertainty*, *1*(1), 7–59. https://doi.org/10.1007/BF00055564

Rogers, H. (1987). *Theory of recursive functions and effective computability*, *Vol. 5*. New York: McGraw-Hill.

Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*, *2*(3), 160. https://doi.org/10.1007/s42979-021-00592-x

Shaffer, V. A., Probst, C. A., 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. https://doi.org/10.1177/0272989X12453501

Thaler, R. (1980). Toward a positive theory of consumer choice. *Journal of Economic Behavior & Organization*, *1*(1), 39–60. https://doi.org/10.1016/0167-2681(80)90051-7

Tversky, A., & Kahneman, D. (1991). Loss Aversion in Riskless Choice: A Reference-Dependent Model. *The Quarterly Journal of Economics*, *106*(4), 1039–1061. https://doi.org/10.2307/2937956

Yang, L., Bian, Y., Zhao, X., Liu, X., & Yao, X. (2021). Drivers' acceptance of mobile navigation applications: An extended technology acceptance model considering drivers' sense of direction, navigation application affinity and distraction perception. *International Journal of Human-Computer Studies*, *145*, 102507. https://doi.org/10.1016/j.ijhcs.2020.102507