

Algorithmic Transference: People Overgeneralize Failures of AI in the Government

Journal of Marketing Research
 2023, Vol. 60(1) 170-188
 © American Marketing Association 2022
 Article reuse guidelines:
sagepub.com/journals-permissions
 DOI: 10.1177/00222437221110139
journals.sagepub.com/home/mrj



Chiara Longoni , Luca Cian , and Ellie J. Kyung 

Abstract

Artificial intelligence (AI) is pervading the government and transforming how public services are provided to consumers across policy areas spanning allocation of government benefits, law enforcement, risk monitoring, and the provision of services. Despite technological improvements, AI systems are fallible and may err. How do consumers respond when learning of AI failures? In 13 preregistered studies (N = 3,724) across a range of policy areas, the authors show that algorithmic failures are generalized more broadly than human failures. This effect is termed “algorithmic transference” as it is an inferential process that generalizes (i.e., transfers) information about one member of a group to another member of that same group. Rather than reflecting generalized algorithm aversion, algorithmic transference is rooted in social categorization: it stems from how people perceive a group of AI systems versus a group of humans. Because AI systems are perceived as more homogeneous than people, failure information about one AI algorithm is transferred to another algorithm to a greater extent than failure information about a person is transferred to another person. Capturing AI’s impact on consumers and societies, these results show how the premature or mismanaged deployment of faulty AI technologies may undermine the very institutions that AI systems are meant to modernize.

Keywords

algorithms, artificial intelligence, social categorization, social impact, government, public policy

Online supplement: <https://doi.org/10.1177/00222437221110139>

Artificial intelligence (AI) is rapidly sweeping through the government and transforming how our core federal and state agencies provide services to consumers. Despite popular belief that the government relies on antiquated procedures, AI decision systems are already fully deployed across many policy areas, including adjudication of government benefits and privileges; the enforcement of the law and regulatory mandates centered on market efficiency, workplace safety, consumer protection, health care, and environmental protection; the monitoring and analysis of risks to public health and safety; and the provision and communication of services to the public. Indeed, a survey by the Administrative Conference of the United States (2020) revealed that nearly half of federal departments, agencies, and subagencies examined have already experimented with AI. The presence of AI in the government is expected to increase further. For instance, the National AI Initiative Act of 2020, which became law in 2021, accelerated AI research and applications across public administrations to foster economic prosperity and national security.

The spread of AI systems across the government promises to reduce costs and improve the quality, consistency, and predictability

of agencies’ decisions, ultimately making public agencies more effective and citizens more satisfied (Administrative Conference of the United States 2020; De Sousa et al. 2019). Unfortunately, despite technological improvements and betterments in performance, AI systems are fallible and may commit errors. For instance, the AI system employed by the state of Arkansas to allocate disability benefits ended up cutting, denying, or terminating caregiver hours without merit and in violation of due process (Lecher 2018). The state of Michigan employed a flawed automated system that incorrectly charged tens of thousands of Michigan residents with fraud and seized millions of dollars in their wages and tax returns (De La Garza 2020). In the last three years alone, the press reported

Chiara Longoni is Assistant Professor of Marketing, Questrom School of Business, Boston University, USA (email: clongoni@bu.edu). Luca Cian is Killgallon Ohio Art Associate Professor of Business Administration, Darden School of Business, University of Virginia, USA (email: cianl@darden.virginia.edu). Ellie J. Kyung is Associate Professor, Marketing Division, Babson College, USA (email: ekyung@babson.edu).

over 370 incidents linked to AI and algorithms, of which 140 pertained to government functions (AIAAIC 2022).

In 2020, Presidential Executive Order 13960 (Executive Office of the President 2020) was issued to regulate the use of AI in society and foster public trust in this technology, and later in the same year, the Consumer Product Safety Commission made AI a top priority in its operating plan. Despite the growing spread of AI systems in our society and therefore its impact on consumers, little is known about consumer responses to AI failures. Most of the research on AI in the government has in fact focused on the technical strengths and weaknesses of AI systems (i.e., what AI systems can or cannot do) and on the imperative to make the use of these systems transparent to consumers (e.g., National AI Initiative Act), rather than on how consumers respond when learning of the faulty deployment of AI systems. The provision of services to consumers is a statutory duty of governments and a source of their legitimacy (Calo and Citron 2021), and therefore examining consumer responses is a critical yet overlooked dimension of assessing the social impact of AI.

Our research fills this gap and seeks to understand consumer responses to failures of AI in the government. We do so by assessing the inferential judgments that consumers make when learning of AI failures, the process that underlies these inferences, and their consequences for propensity to apply for public services and trust in the government. In 13 preregistered studies (combined $N=3,724$) across several policy areas, we show that algorithmic failures are generalized more broadly than human failures. We term this effect “algorithmic transference,” as it is an inferential process that generalizes (i.e., transfers) information about one member of a group to another member of that same group. Rather than reflecting generalized algorithm aversion (Dietvorst, Simmons, and Massey 2015), algorithmic transference is due to group categorization processes: it stems from how people perceive a group of nonhuman agents compared with a group of human agents. AI systems are viewed as out-groups characterized by greater homogeneity than in-groups of humans. Because AI algorithms are viewed as a highly homogeneous group, information learned about one algorithm is transferred to another algorithm to a greater extent than information learned about a member of a more heterogeneous group—a person. We provide evidence of perceived group homogeneity as a driver of algorithmic transference through mediation and moderation, and by delineating the scope of the effect. Finally, we show how algorithmic transference may have detrimental consequences for propensity to access public services, whose provision is a core duty of the government.

Our research and findings make the following theoretical contributions. We contribute to research on consumer responses to algorithmic failures by identifying the novel effect of algorithmic transference: greater generalization of algorithmic than human failure information. This effect is novel, as prior research has focused either on changes in reliance on an algorithm (benchmarked on changes in reliance on a person; cf. Berger et al. 2021; Dietvorst, Simmons, and Massey 2015;

Prahl and Van Swol 2017) or on moral judgments ensuing from learning of algorithmic (vs. human) failures (Awad et al. 2020; Gill 2020; Srinivasan and Sarial-Abi 2021). Second, we contribute to the broader literature on responses to AI systems (e.g., Cadario, Longoni, and Morewedge 2021; Castelo, Bos, and Lehmann 2019; Longoni and Cian 2020) by identifying a novel psychological process—perceptions of group homogeneity—that shapes consumer responses to AI systems. Whereas prior theoretical accounts have implicated lay beliefs about an individual algorithmic agent (e.g., incapacity to learn, lack of intentionality), we propose a novel perspective based on social categorization and on how people view algorithmic agents at the group level (i.e., as more homogeneous than a group of people).

Our research is also novel from the substantive perspective of assessing AI’s social impact. Gauging AI’s impact on societies cannot be limited to the technological aspects related to which AI systems are developed. The social impact of AI also needs to encompass the psychological aspects related to how consumers respond to AI’s deployment. The deployment of AI may enable the government to meet its statutory duties toward consumers and deliver services more effectively (Calo and Citron 2021). However, as we highlight in this research, the premature fielding of faulty AI technologies may engender algorithmic transference and dampen consumer propensity to utilize public services and trust in the government, ultimately undermining the very institutions that AI systems are meant to modernize.

Theoretical Development

Responses to AI Failures

The investigation of people’s propensity to rely on AI, algorithms, and other forms of automation has come to the fore in recent years given the increasing pervasiveness of these systems in our lives. Research in this area has documented aversion to relying on automated systems over humans due to lay beliefs that these systems are unfit for subjective, hedonic, or personally unique tasks (Castelo, Bos, and Lehmann 2019; Longoni, Bonezzi, and Morewedge 2019, 2020; Longoni and Cian 2020); that they undermine self-expression (Granulo, Fuchs, and Puntoni 2020); or that they are black boxes (Cadario, Longoni, and Morewedge 2021). A parallel stream of research has documented appreciation of automated systems over humans under certain circumstances: if concerns about inequality or unfair treatment are salient (Bigman et al. 2021), if the task is impersonal and people are unfamiliar with the system (Berger et al. 2021), if objectivity (Castelo, Bos, and Lehmann 2019) or utilitarian goals are salient (Longoni and Cian 2020), or if quantitative estimates are made (Logg, Minson, and Moore 2019).

Within this literature, two streams are particularly relevant for our research as they examine responses to AI failures. One stream of research has focused on changes in reliance on (i.e., stated or revealed preference for) an algorithm or a

human before and after learning of failures of that same algorithm/human. This research shows that people are indifferent between an algorithmic and a human advisor when their respective performances are unknown (Dietvorst, Simmons, and Massey 2015) or when people are unfamiliar with the advisor (Berger et al. 2021). After observing an advisor err or dispense bad advice, however, reliance on the algorithm decreases more than reliance on the human (Dietvorst, Simmons, and Massey 2015; Dzinzolet et al. 2002; Prah and Van Swol 2017) unless the algorithm has the capacity to learn (Berger et al. 2021).

Another stream of research has focused on moral judgments ensuing from AI failures. In the domain of brand scandals, people are more forgiving and respond less negatively when brand harms are caused by algorithmic than human failures, an effect due to lower attribution of the harm to algorithms than to humans (Srinivasan and Sarial-Abi 2021). These findings are echoed by research on responses to autonomous vehicles: people view harm to pedestrians by an automated vehicle as more permissible than harm caused by themselves or a regular driver (Awad et al. 2020; Gill 2020). Indeed, people are less morally outraged by algorithmic discrimination than by human discrimination (Bigman et al. 2021).

Our theoretical and substantive focus departs from prior research in the following ways. Unlike prior research, we focus neither on reliance nor on moral judgment of a failing agent. Stated and revealed preference are often inadequate to capture responses in the public sector, a domain where consumers cannot typically choose whether a person or an AI system will be their service provider. For instance, consumers cannot choose whether their unemployment benefits will be allocated by a person or an automated decision system, or whether their insurance claims will be checked for fraud by a person or an algorithm. However, consumers can choose whether to apply to use such public services at all. Thus, we focus on inferential judgments—the degree of transfer of information from one member of a group to another member of that same group—and their implications for propensity to apply for public services. That is, whereas prior research examined judgments of the *same* algorithm or human after learning failure information (i.e., judgment of algorithm A or human A before and after failure information; e.g., Berger et al. 2021; Dietvorst, Simmons, and Massey 2015), we investigate (1) the extent to which failure information about an algorithm or a human is transferred to a *different* algorithm or human (i.e., from algorithm A to algorithm B or from human A to human B), and (2) the consequences for consumer propensity to utilize public services. By focusing on inferential judgments and their downstream consequences, our research captures how consumers might respond to the reporting of AI failures by the news media.

Social Categorization and Algorithmic Transference

We situate our predictions within the framework of social identity theory (Tajfel 1969), which focuses on how group membership influences both one's self-concept and one's relations with in-group and out-group members. At the foundation of social identity theory is self-categorization, which describes how

people categorize themselves and others into different social groups (Aydinoğlu and Cian 2014; Hogg and Reid 2006; Pandya, Cian, and Venkatesan 2022; Turner 1982). Based on accessible social identities, people assign themselves and others to social categories, thus distinguishing between in-groups, groups in which people are members, and out-groups, groups to which people do not belong.

We expect these social categorization processes to apply to how people perceive human versus nonhuman agents. Specifically, we assume that people view nonhuman agents, such as AI systems, as social entities, and categorize them as out-groups—members of a group people do not belong to. This assumption is consistent with research in marketing, psychology, and human–computer interaction, showing that people ascribe social categories to autonomous artificial agents and that these social categories then shape interactions with artificial agents. For instance, people apply gender (Borau et al. 2021; Nass, Moon, and Green 1997) and occupation (Tay, Jung, and Park 2014) stereotypes to automated agents, reciprocate acts by helpful computers (Fogg and Nass 1997), and rely on a host of social cues when interacting with artificial agents (Nass and Moon 2000; Reeves and Nass 1996). Notably, both embodied and virtual artificial agents seem to trigger application of social categories (Rossen et al. 2008), so much so that according to the computers-as-social-actors paradigm, people mindlessly apply social categories to computers even though they know that computers lack feelings and human motivations (Nass and Moon 2000; Nass, Steuer, and Tauber 1994).

An empirical illustration further validated our assumption that people view AI systems as social agents and, more specifically, as out-groups. We asked respondents from Amazon Mechanical Turk (MTurk; $N = 300$; $M_{\text{age}} = 40.1$ years, $SD = 11.8$; 46.3% female, 53.3% male, .3% prefer not to say) to read brief descriptions of three human agents and three algorithmic agents performing various tasks. We used the same descriptions employed in the studies that constitute the empirical package (e.g., “Consider an algorithm employed by an agency of the federal government to calculate unemployment benefits to distribute to citizens”; “Consider a person employed by an agency of the federal government to review consumer complaints, identify trends, and predict consumer harm in textual consumer complaints”). To test our prediction that people view algorithms as out-groups and other humans as in-groups, we asked participants to rate the extent to which they viewed each target [algorithmic/human] agent as part of a group that they belonged to as well, a member of their same group, and belonging to the same group as they did (1 = “Not at all,” and 7 = “Very much”; statements were presented in random order and based on Ostrom and Sedikides [1992]; $\alpha = .78$). Thus, lower numbers indicate greater categorization as out-groups and higher numbers indicate greater categorization as in-groups. Validating our prediction, participants viewed algorithms as out-groups ($M_{\text{AI}} = 1.86$, $SD = 1.41$) to a greater extent than they viewed humans as out-groups ($M_{\text{H}} = 5.34$, $SD = 1.56$; $t(290) = 28.99$, $p < .001$, $d = 1.7$ [nine missing values]; details in Web Appendix A).

Building on this assumption, we propose that categorizing AI algorithms as out-groups results in perceiving them as a more homogeneous and undifferentiated group than humans (i.e., “all AI algorithms are the same”). Indeed, the principle of optimal distinctiveness underlying social categorization leads to accentuating intergroup differences and intragroup similarity (Tajfel 1969), an effect called out-group homogeneity—the tendency to perceive out-groups as less variable and more homogeneous than members of one’s own group (in-groups) on several attributes (Linville and Fischer 1998; Ostrom and Sedikides 1992; Sedikides and Ostrom 1993). Out-group homogeneity has been documented across many naturally occurring groups, such as race, religion, nationality, and profession (Brewer 1993; Park, Judd, and Ryan 1991).

We further predict that because AI algorithms are perceived as a group having higher homogeneity than a group of people, information about one algorithm is generalized and transferred to another algorithm to a greater extent than information about one person is generalized and transferred to another person. This prediction is supported by research suggesting that members of homogeneous groups may be viewed as exemplifying the group’s central tendency, and therefore information about an individual member of a homogeneous group is more likely to be generalized to the whole group (Nisbett et al. 1983; Quattrone and Jones 1980). Because AI systems are perceived as a more homogeneous group than people, failure information about one AI system is transferred to another AI system to a greater extent than failure information about one person is transferred to another person.

In summary, our key prediction is that people generalize algorithmic failures to a greater extent than human failures—an effect we term “algorithmic transference” and assess via inferential judgment of failure. We further predict that algorithmic transference is due to social categorization processes, which we assess by measuring perceptions of homogeneity within a group (of humans or AI systems). As out-groups, AI systems are viewed as a group of higher homogeneity than people, who are in-groups. Higher homogeneity results in greater transference of failures for AI systems than people.

Two final considerations warrant mention. First, people’s perception of AI algorithms as an undifferentiated group is not necessarily grounded in practice. Because AI systems are either contracted to various third parties or developed in-house by each governmental agency (Administrative Conference of the United States 2020), it is unlikely that these systems are similar to each other. Second, our predictions with respect to algorithmic transference are rooted in perceptual and representational processes that need not implicate nor oppose algorithm aversion. That is, algorithmic transference is distinct and may operate independently from generalized algorithm aversion, as explicated in more detail in Study 3.

Overview of the Studies

We systematically examined consumer responses to algorithmic failures (both biases and errors in decisions) in a series of

preregistered studies across a range of policy areas (allocation of disability, social security, unemployment, and Temporary Assistance for Needy Families [TANF] government benefits; fraud determinations; consumer protection services) and diverse samples (convenience, U.S. nationally representative, experts). To maximize external validity, we selected policy areas where AI systems are in use by relying on a report prepared for the Administrative Conference of the United States (2020).

Studies 1a–1c and the replication reported in Web Appendix B tested algorithmic transference by using real news articles. Studies 2 and 3 tested the proposed process account: group homogeneity perceptions. Study 2 provided process evidence via mediation, testing group homogeneity perceptions along with possible alternative explanations (general knowledge of AI and algorithms, perceived locus of causality) while also controlling for base rates. Study 3 provided process evidence via moderation, testing for the elimination of transference by making group heterogeneity salient. Studies 4 and 5 explored the scope of algorithmic transference, testing whether the effect varies with a person’s degree of discomfort with new technologies and is eliminated with human oversight. Studies 6a–6c investigated the implications for consumers and policy, testing the consequences of algorithmic transference for propensity to utilize public services. Ancillary Studies 7 and 8, reported in the “General Discussion” section, examined the generalization of brand transference to brand scandals and the implications for trust in the government. Figure 1 provides an overview of the conceptual framework and studies.

All the studies reported in the article are preregistered. We report all conditions, data exclusions, and measures collected. We report the experimental stimuli for all studies in Web Appendix A and the preregistrations links in Web Appendix C. Our target preregistered samples were 100 per cell in each study with the following exceptions: In Studies 1b and 6a, we collected 150 responses per cell as that was the minimum number of responses required by the platform Prolific to yield a nationally representative U.S. sample. In Study 2, we collected 2.5 times the number of observations per condition (250 instead of 100) because we were testing a potential interaction with a continuous variable. In all the studies and as preregistered, we programmed the Qualtrics surveys to automatically exclude participants who failed an attention check (i.e., a simple arithmetic calculation) at the very beginning of the survey and prior to any manipulation. We did not collect data for these participants, and they did not affect our target sample size.

Studies 1a–1c: Evidence of Algorithmic Transference

The first set of studies tested for algorithmic transference—greater generalization of algorithmic than human errors—through information that would mimic the way in which people learn about algorithmic failures as much as possible: by reading an article in the news or via social media. Participants read real

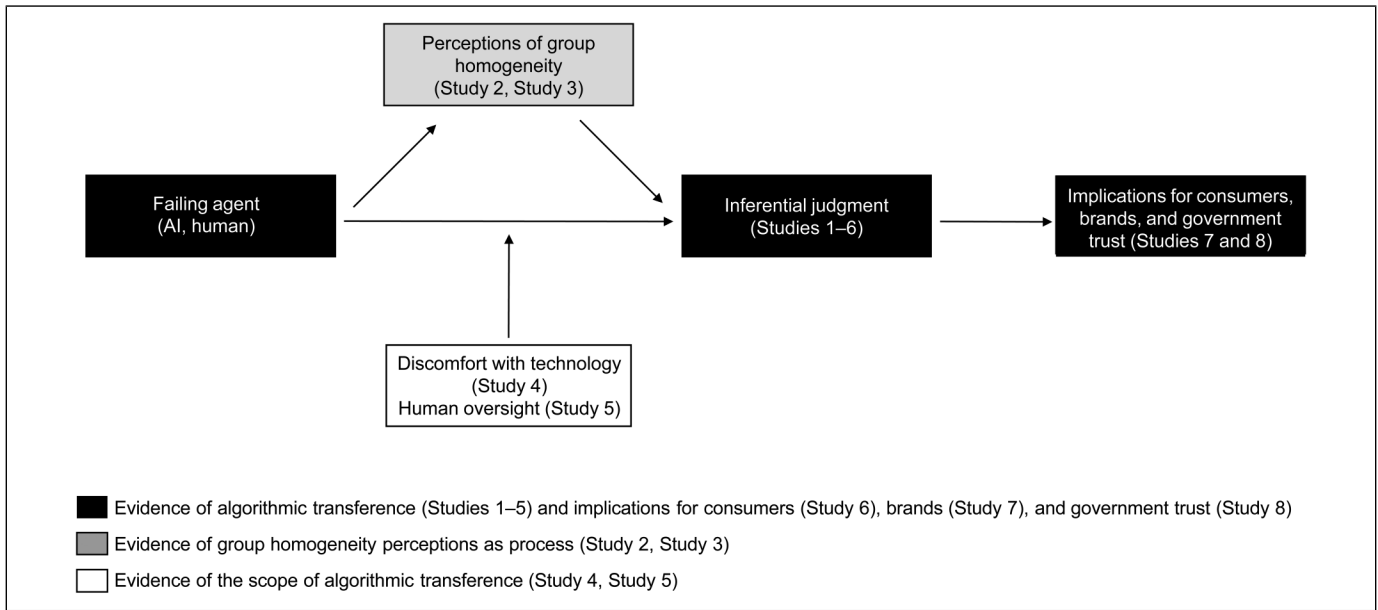


Figure 1. Conceptual Framework and Overview of the Studies.

news articles describing either an algorithmic or a human failure in the allocation of disability benefits (Study 1a), calculation of social security benefits (Study 1b), or determination of fraud in insurance claims (Study 1c). We conducted these studies on a convenience sample (Study 1a), a nationally representative U.S. sample (Study 1b), and a sample of technology experts (Study 1c).

Procedure

A total of 707 respondents participated in exchange for \$.35 (Study 1a: $N = 205$; $M_{\text{age}} = 38.1$ years, $SD = 10.9$; 44.0% female, 55.0% male, 1.0% nonbinary/third gender; Study 1b: $N = 300$, $M_{\text{age}} = 44.3$ years, $SD = 16.0$, 51.3% female, 47.7% male, 1.0% nonbinary/third gender; Study 1c: $N = 202$, $M_{\text{age}} = 41.3$ years, $SD = 11.7$, 24.0% female, 74.5% male, 1.5% nonbinary/third gender).

Study 1a: Allocation of disability benefits. We recruited respondents from MTurk. In a two-cell, between-subjects design, participants read a news article describing the failure of either an algorithm or a person employed by the state of Arkansas to allocate caregiver benefits to the state's residents. This article was based on a piece by Lecher (2018) that appeared in *The Verge*.

What Happens When [An Algorithm/A Person] Cuts Your Healthcare

Tammy Dobbs moved to the state of Arkansas in 2008 and signed up for a state disability program to help her with her cerebral palsy. Under the program, the state had to determine the number of caregiver hours she would need. Because Tammy spent most of her waking hours in a wheelchair and had stiffness in her hands, she was allocated 56 hours of home care per week. In 2016, the state of Arkansas employed [an algorithm/a person] to recalculate the number of caregiver hours Tammy would be allotted. Without

any explanation or opportunity for comment, discussion, or reassessment, the [algorithm/person] allotted Tammy 32 hours per week, a massive and sudden drop that Tammy had no chance to prepare for and that severely reduced her quality of life.

After reading the article, participants made an inferential judgment about the likelihood that another agent of the same group (i.e., an algorithm or a person) employed by a different state (Kentucky) would fail. Specifically, participants rated the likelihood that [an algorithm/a person] employed by the state of Kentucky would make wrong disability benefits calculations (1 = "Unlikely to make wrong calculations," and 7 = "Likely to make wrong calculations"). In this study and in the subsequent studies, we did not recruit respondents from the states mentioned in the survey (in this case, Arkansas and Kentucky). These respondents were explicitly excluded from our recruitment instructions and automatically prevented from participating by the survey software (Qualtrics), and we have no data for them. We implemented this preregistered exclusion to avoid a potential source of confusion. As we used real articles, we wanted to avoid cases where a participant assigned to the human condition and residing in the state mentioned in the article was aware that the failure reported in the article involved an algorithm and not a person. State of residence does not, however, affect algorithmic transference, as evidenced by an ancillary study reported in Web Appendix B in which we replicated the results of Study 1a recruiting participants only from the state mentioned in the stimuli. Finally, in this and all the other studies, the survey ended with a manipulation check (whether participants correctly recalled the agent described: an algorithm or a person) and questions capturing demographic variables.

Study 1b: Calculation of social security benefits. We recruited respondents from a nationally representative U.S. sample

from Prolific, which uses quota sampling to provide a sample that is matched to the U.S. population on sex, age, and ethnicity. In a two-cell, between-subjects design, participants read a news article describing the failure of either an algorithm or a person employed by the state of Michigan to calculate unemployment insurance benefits. This article was based on a piece by Goodwin (2020) that appeared in *The London Economic*.

Poorly trained Universal Credit [algorithm/person] forces people into hunger and debt. A new report found that the government benefit system—which is reliant on [an algorithm that/a person who] calculates benefits for people on a low income or out of work—threatens the rights of people most at risk of poverty.

People in the UK are being pushed into poverty by [a computer algorithm that/a person who] incorrectly allocated how much money they are entitled to in social security payments. Universal Credit claimants are being forced to forego food and take on debt because of a poorly trained algorithm, a leading human rights charity warned. A new report by Human Rights Watch found that the government benefit system—which is reliant on [an algorithm/a person] to calculate benefits for people on a low income or out of work—“threatens the rights of people most at risk of poverty”. The charity is calling on the government to take action ahead of an anticipated winter jobs crisis, with up to nine million workers at risk of redundancy as the furlough scheme is wound down.

After reading the article, participants read information about the Social Security program in the United States (“The Social Security program in the United States provides protection against the loss of earnings due to retirement, death, or disability”) and made an inferential judgment about the likelihood that [an algorithm/a person] employed by the state of Arkansas would make wrong Social Security benefits calculations (1 = “Unlikely to make wrong calculations,” and 7 = “Likely to make wrong calculations”).

Study 1c: Determination of unemployment insurance fraud. We recruited respondents from Cloud Research, selecting those who indicated “computer science/engineering” and “high-tech” as their field of training or current work. This sample allowed us to test whether algorithmic transference manifested when a certain degree of knowledge with technology may be presumed.

In a two-cell, between-subjects design, participants read a news article describing the failure of either an algorithm or a person employed by the state of Michigan to determine fraud in unemployment insurance claims. This article was based on a piece by De La Garza (2020) that appeared in *Time*.

Michigan Fraud Detection [Algorithm/Person] Trapped Citizens in Bureaucratic Nightmares with Their Lives on the Line

Lindsay Perry was 30 weeks pregnant and on bedrest when her husband Justin was accused by a Michigan [automated system/person] of unemployment fraud and fined \$10,000 after losing his job as a chef. Their tax returns were seized for three years in a row, their van was repossessed, and they filed for bankruptcy.

Later, Michigan reversed the charges as the fraud determination was incorrect and reimbursed the couple \$6,000, but the damage was already done. Perry’s husband was one of many people across Michigan who were wrongly accused of unemployment insurance fraud as a result of a poorly trained [algorithmic system operated/person employed] by the state.

Then, participants made an inferential judgment about the likelihood that [an algorithm/a person] employed by the state where they lived would erroneously check for unemployment benefit fraud (1 = “Unlikely to erroneously check for fraud,” and 7 = “Likely to erroneously check for fraud”).

Results and Discussion

As we predicted, participants were more prone to transfer algorithmic than human failures. Across all three studies, participants were more prone to infer that another algorithm, but less so another person, would fail (Study 1a: $M_{AI} = 5.32$, $SD = 1.40$; $M_H = 4.84$, $SD = 1.23$; $t(203) = 2.60$, $p = .01$, $d = .36$; Study 1b: $M_{AI} = 4.81$, $SD = 1.52$; $M_H = 4.30$, $SD = 1.47$; $t(298) = 3.01$, $p = .003$, $d = .35$; Study 1c: $M_{AI} = 5.30$, $SD = 1.44$; $M_H = 4.78$, $SD = 1.59$; $t(200) = 2.41$, $p = .017$, $d = .34$).

Together, this first set of studies showed evidence of algorithmic transference—greater inferential generalizations of algorithmic than human errors. The effect was robust across policy areas (calculation of disability and social security benefits, fraud determination), across populations (a convenience sample, a representative U.S. sample), and when employing a sample of experts with a presumably moderate degree of knowledge of technology. The next study tested whether perceived group homogeneity perceptions drive algorithmic transference relative to alternative explanations and while controlling for base rate information.

Study 2: Testing Perceptions of Group Homogeneity as Process

Study 2 tested perceptions of group homogeneity as a driver of algorithmic transference along with two potential alternative explanations: general knowledge of AI algorithms and perceived locus of causality. Our prediction was that people perceive AI algorithms as a group of higher homogeneity than humans, and that differential group homogeneity perceptions explain algorithmic transference. One potential alternative explanation is based on lack of general knowledge and understanding of AI algorithms. People may have a low degree of knowledge and a poor understanding of AI algorithms, and low algorithmic literacy may result in greater transference. We addressed this explanation indirectly in Study 1c, by recruiting a sample of experts, and we tested it more directly in Study 2 by measuring algorithmic literacy. If algorithmic literacy explains transference, we should observe greater transference the less a person knows about AI and algorithms (i.e., there should be an interaction between algorithmic literacy and the type of agent). Another potential alternative explanation is

based on locus of causality of the failure (internal to the agent versus external; Ryan and Connell 1989). Whereas people may attribute a human failure to external causes less likely to apply to other humans (e.g., “she was having a bad day”), people may attribute an algorithmic failure to more stable causes internal to the algorithm, which are more likely to apply to other algorithms. Thus, in Study 2 we measured locus of causality and pitted it against our group homogeneity account. Finally, we controlled for base rates in the accuracy of the focal decision. That is, we provided respondents with a reference point for the failure rate of both an algorithmic and a human agent in the domain described to ensure that differential expectations about failure rates between algorithms and human agents could not account for transference.

Procedure

Respondents from MTurk participated in exchange for \$1 ($N = 502$; $M_{\text{age}} = 40.8$ years, $SD = 12.1$; 50.4% female, 47.6% male, .4% nonbinary/third gender, 1.6% prefer not to say).

First, participants read information about the federal–state unemployment insurance program, which contained base rate information about the accuracy with which unemployment insurance benefits are calculated:

The nation’s unemployment insurance program is a federal-state system that provides temporary income support for unemployed workers. The system is funded by taxes collected from employers and held in trust funds administered by individual states. States—which are charged with distributing and overseeing many federally funded benefits—are taking these calculations seriously. On average, between 35 and 40 percent of unemployment benefits calculations are inaccurate.

Then, in a two-cell, between-subjects design, participants read a news article describing a failure by either an algorithm or a person employed by the state of Michigan to calculate unemployment benefits.

In the recent past, the state of Michigan employed [an algorithm/a person] to calculate unemployment benefits. The state then allocated unemployment benefits to its residents based on the [algorithm’s/person’s] calculations. As it turns out, a state review later determined that most of these calculations completed by the Michigan [algorithm/person] were wrong.

Participants then made an inferential judgment about the likelihood that [an algorithm/a person] employed by the state of Arkansas to calculate unemployment benefits would make wrong calculations (1 = “Unlikely to make wrong calculations,” and 7 = “Likely to make wrong calculations”). To measure group homogeneity perceptions, participants were first prompted to think about the [algorithms/people] performing tasks like the one described in the article in various agencies and states across the country. Then, participants rated the extent to which they agreed with the following statements: “These [algorithms/people] are most likely very similar to each other

in terms of their characteristics,” “These [algorithms/people] most likely resemble one another in terms of their characteristics and capabilities,” and “These [algorithms/people] most likely share common underlying characteristics” (1 = “Disagree strongly,” and 7 = “Agree strongly”; items presented in random order; $\alpha = .95$). To measure perceived locus of causality, participants rated the extent to which they thought that the [algorithm/person] carried out the task described because of intrinsic motivation versus extrinsic motivation, causes within the [algorithm/person] versus causes external to the [algorithm/person], internal reasons versus external reasons, and the task’s own sake versus some external reason (based on Ryan and Connell [1989]; items were measured on seven-point bipolar scales and presented in random order; $\alpha = .92$).

Then, participants completed a test of algorithmic literacy intended to measure their general knowledge and understanding of AI and algorithms. First, they read a quick introduction explaining the goal of the task:

In the next pages we will ask you 16 questions about your general knowledge of Artificial Intelligence and algorithms. It is very important that you do not look up or google these answers online, as that would defeat the purpose of our survey. We are interested in people’s actual understanding of Artificial Intelligence and algorithms. Please try to answer these questions to the best of your knowledge. We need your honest responses to understand people’s general knowledge of AI and algorithms.

Then, participants completed the algorithmic literacy test. We developed the algorithmic literacy test by compiling an initial battery of 20 questions and then asked an academic and a practitioner, both experts on AI and algorithms, to validate these questions (i.e., check for accuracy in verbiage and answers). From the experts’ feedback, we eliminated four questions and refined the verbiage of all questions. The final test comprised 16 multiple-choice questions: 8 questions about AI and 8 questions about algorithms. Each question had three possible answers, only one of which was correct. The order of questions and the order of answers within each question were randomized. We report the final algorithmic literacy test in Web Appendix D.

Results and Discussion

We computed an algorithmic literacy score for each participant by assigning a score of 1 to each correct answer and 0 to each incorrect answer and summing the number of correct responses. Each participant’s algorithmic literacy score thus ranged between 0 and 16 ($M = 9.89$, $SD = 2.47$, median = 10.00). The distribution of correct responses by question and across participants is reported in Web Appendix D (skewness: $-.122$; kurtosis: $-.208$).

Algorithmic transference. We first tested whether algorithmic literacy moderated algorithmic transference. We regressed inferential judgment on agent (dummy coded: AI = 1, human = 0), algorithmic literacy score ($M = 9.89$, $SD = 2.47$), and the interaction between agent and algorithmic literacy score (mean-

centered). The analysis revealed a significant main effect of agent ($\beta = .267$, $t(498) = 6.15$, $p < .001$), no significant effect of algorithmic literacy score ($\beta = -.065$, $t(498) = -1.11$, $p = .268$), and no significant two-way interaction between agent and algorithmic literacy score ($\beta = .020$, $t(498) = .34$, $p = .736$), suggesting that algorithmic literacy does not moderate transference. Following up on the significant main effect of agent, a t-test on inferential judgment replicated algorithmic transference: participants were more prone to generalize algorithmic failures than human failures, inferring that another algorithm, more so than another person, would fail ($M_{AI} = 5.44$, $SD = 1.18$; $M_H = 4.77$, $SD = 1.25$; $t(500) = 6.13$, $p < .001$, $d = .55$).

Group homogeneity perceptions. We then tested whether algorithmic literacy moderated group homogeneity perceptions. We regressed group homogeneity perceptions on agent (dummy coded: AI = 1, human = 0), algorithmic literacy score, and the interaction between agent and algorithmic literacy score (mean-centered). The analysis revealed a significant main effect of agent ($\beta = .426$, $t(498) = 10.60$, $p < .001$), a marginal effect of algorithmic literacy score ($\beta = .106$, $t(498) = 1.94$, $p = .053$), and no significant two-way interaction between agent and algorithmic literacy score ($\beta = .002$, $t(498) = .039$, $p = .969$), suggesting that algorithmic literacy does not moderate group homogeneity perceptions. Following up on the significant main effect of agent, a t-test on group homogeneity perceptions supported our prediction that participants viewed algorithms as part of a more homogeneous group than people ($M_{AI} = 5.60$, $SD = 1.05$; $M_H = 4.46$, $SD = 1.31$; $t(500) = 10.71$, $p < .001$, $d = .96$).

Perceived locus of causality. A t-test on perceived locus of causality index with agent as independent variable was significant: participants attributed the algorithmic decision to internal causes to a greater extent than the human decision ($M_{AI} = 4.17$, $SD = 1.96$; $M_H = 5.10$, $SD = 1.36$; $t(499) = 6.18$, $p < .001$, $d = -.55$ [one missing value]).

Mediation. We then conducted a mediation analysis to test whether group homogeneity perceptions mediated the observed differences in algorithmic transference. We tested a mediation model that included, along with group homogeneity perceptions, perceived locus of causality as a simultaneous mediator. We examined confidence intervals (CIs) using 10,000 bootstrap iterations (Hayes 2018, PROCESS Model 4) coding agent as 1 when AI and 0 when human. The indirect effect of agent on inferential judgment through group homogeneity perceptions was significant (.48, 95% CI: [.34, .63]), and the direction of the effects confirmed that an algorithm was associated with greater group homogeneity perceptions than a person, which in turn contributed to greater algorithmic transference. When controlling for this indirect effect, the direct effect of agent on inferential judgment was no longer significant (.19, 95% CI: [-.03, .42]; total effect: .67, 95% CI: [.45, .88]). Whereas the indirect effect through group homogeneity perceptions was significant, the indirect effect through perceived locus of causality was not (-.01,

95% CI: [-.06, .05]), indicating that this variable did not account for transference (details in Web Appendix E).

These results provided evidence of group homogeneity perceptions as a driver of algorithmic transference. Furthermore, we did not find evidence that transference of algorithmic failures was driven by perceived locus of causality. Finally, we did not find evidence supporting the notion that algorithmic transference was moderated by algorithmic literacy. These results suggest that algorithmic transference is neither due to differential lack of knowledge and understanding of AI and algorithms nor to attribution of the failure to external/internal causes. In Study 3 we tested group homogeneity as driver of transference via moderation.

Study 3: Making Out-Group Heterogeneity Salient Eliminates Algorithmic Transference

Study 3 tested a theoretically driven and practically relevant moderator that should curb algorithmic transference: making out-group heterogeneity salient. If transference is due to perceptions of algorithms as a homogeneous group, interventions that dispel this belief should eliminate transference. We made out-group heterogeneity salient by specifying that the failing agent belonged to a group made up of members that were likely to share few similarities, were unlikely to resemble one another in terms of their characteristics and capabilities, and were not a homogeneous group. This intervention served to make group differences salient and thus have AI systems be perceived as a more varied, heterogeneous group, in line with research on intergroup perception showing how third-party communication about the characteristic level of homogeneity of a group shapes people's stored beliefs about the variability of a group (Ostrom and Sedikides 1992; Park and Hastie 1987).

This study also allowed us to test a social categorization account based on group representational processes (our group homogeneity account) and an account based on algorithm aversion. An algorithm aversion account, for instance due to perceptions that algorithms are unfit to carry out the focal task or due to a generalized resistance to automated systems, would predict that algorithmic transference will manifest both in the control condition and in the condition where group heterogeneity is salient.

Finally, this study allowed us to test our group homogeneity account against an explanation based on higher standards for algorithms than humans. That is, consumers may hold lay theories that algorithms are not supposed to fail, and thus react more negatively when learning of algorithmic failures than human failures. This alternative explanation would also predict higher failure likelihood for an algorithm than for a person irrespective of whether algorithms are described as part of a heterogeneous group or not.

Procedure

Respondents from MTurk participated in exchange for \$.35 ($N = 403$; $M_{age} = 39.7$ years, $SD = 11.5$; 43.2% female, 55.3% male, .5% nonbinary/third gender, 1.0% prefer not to say).

First, participants read information about federal unemployment benefits distributed by states in the United States. As in Study 2, this information contained base rate information about the accuracy with which unemployment insurance benefits are calculated:

The nation's unemployment insurance program is a federal-state system that provides temporary income support for unemployed workers. The system is funded by taxes collected from employers and held in trust funds administered by individual states. States—which are charged with distributing and overseeing many federally funded benefits—are taking these calculations seriously. On average, between 35 and 40 percent of unemployment benefits calculations are inaccurate.

Participants were then randomly assigned to one condition in a 2 (agent: algorithm, human) \times 2 (heterogeneity: salient, control) between-subjects design. We used the same scenario as in Study 2 with minor wording differences given the different factors manipulated. Specifically, to manipulate the agent, participants read a news article describing a failure in the calculation of unemployment benefits by either an algorithm or a person employed by the state of Michigan:

In the recent past, the state of Michigan employed [an algorithm/a person] to calculate unemployment benefits. The state then allocated unemployment benefits to its residents based on the [algorithm's/person's] calculations. As it turns out, a state review later determined that most of these calculations completed by the Michigan [algorithm/person] were wrong.

Between subjects, and orthogonal to agent, we manipulated whether the group of algorithms/people making the focal decision was described as a heterogeneous group (heterogeneity salient condition) or omitted this specification (control condition). Specifically, participants in the heterogeneity salient condition read:

Note that different states rely on different contractors that train algorithms for these purposes, and therefore the algorithms making these decisions share very few similarities and are unlikely to resemble one another in terms of their characteristics and capabilities, so much so that it would be hard to group these algorithms in one homogeneous group.¹

Participants then made an inferential judgment about the likelihood that [an algorithm/a person] employed by the state of Arkansas would make wrong unemployment benefits calculations (1 = “Unlikely to make wrong calculations,” and 7 = “Likely to make wrong calculations”).

¹ A posttest ($N = 200$, MTurk) confirmed that the heterogeneity manipulation successfully affected perceptions of group homogeneity: participants viewed the group of algorithms in the heterogeneity salient manipulation as less homogeneous ($M = 2.63$, $SD = 1.66$) than the group of algorithms in the control condition ($M = 5.34$, $SD = 1.19$; $t(198) = 13.30$, $p < .001$, $d = 1.88$).

Results and Discussion

A 2 \times 2 analysis of variance on inferential judgment revealed a significant main effect of agent ($M_{AI} = 4.96$, $SD = 1.32$; $M_H = 4.42$, $SD = 1.47$; $F(1, 399) = 15.61$, $p < .001$, $\eta_p^2 = .04$), a significant main effect of heterogeneity ($M_{control} = 4.84$, $SD = 1.42$; $M_{heterogeneity\ salient} = 4.54$, $SD = 1.41$; $F(1, 399) = 4.43$, $p = .036$, $\eta_p^2 = .01$), and a significant two-way agent \times heterogeneity interaction ($F(1, 399) = 6.30$, $p = .012$, $\eta_p^2 = .02$). Planned contrast in the control conditions replicated algorithmic transference ($M_{AI, control} = 5.28$, $SD = 1.23$; $M_{H, control} = 4.39$, $SD = 1.46$; $F(1, 399) = 20.82$, $p < .001$). As predicted, however, algorithmic transference was eliminated when group heterogeneity was salient: participants inferred that an algorithm and a person had the same likelihood to fail ($M_{AI, heterogeneity\ salient} = 4.64$, $SD = 1.32$ vs. $M_{H, heterogeneity\ salient} = 4.45$, $SD = 1.49$; $F(1, 399) = 1.04$, $p = .308$). Further underscoring this point, transference for the algorithm described as belonging to a heterogeneous group was the same as transference for the human in the control condition ($F(1, 399) = 1.70$, $p = .193$). These results are unresponsive of accounts based on algorithm aversion and higher standards for algorithms: if transference was due to generalized negative responses or higher standards for algorithms, participants should have predicted higher failure likelihood for an algorithm than for a person regardless of whether algorithms were described as a heterogeneous group or not. Inferential judgment for the person was unaffected by the group heterogeneity manipulation ($F(1, 399) = .08$, $p = .78$), suggesting that participants represented humans as a heterogeneous group. Finally, inferential judgment for the algorithm was higher when group heterogeneity was salient compared with the control ($F(1, 399) = 10.67$, $p = .001$).

These results corroborated our predictions with respect to the role of group homogeneity perceptions in explaining algorithmic transference. In the control condition, participants were more prone to transfer information about algorithms than about humans, consistent with prior results. When algorithms were specified to be a highly heterogeneous group, however, transference did not manifest. These results underscore another important theoretical point: algorithmic transference is a perceptual process rooted in how people mentally represent algorithms, and not merely an instantiation of algorithm aversion. The next two studies tested the scope of algorithmic transference.

Study 4: Discomfort with Technology Moderates Algorithmic Transference

Study 4 served as a test of the scope of algorithmic transference. Specifically, we examined whether transference is moderated by consumers' degree of discomfort with new technologies. We relied on a measure called the Technology Readiness Index (TRI), which assesses people's general “propensity to embrace and use new technologies for accomplishing goals in home life and at work” (Parasuraman 2000, p. 308). Our focus was on factors that might inhibit consumers' propensity to embrace new technology, and therefore we focused on the subset of the TRI measuring discomfort with new technology.

This subscale captures consumer beliefs, attitudes, and motivations that may curtail adoption of new technologies. In this respect, we expected discomfort with new technology to be positively correlated with transference, as one may be more likely to generalize algorithmic failures the more one negatively views the employment of technology in society.

Procedure

Respondents from MTurk participated in exchange for \$.40 ($N = 400$, $M_{\text{age}} = 41.0$ years, $SD = 11.9$; 49.0% female, 50.5% male, .5% prefer not to say).

Participants began the study by filling out an abbreviated version of the discomfort with technology subscale of the TRI (Parasuraman 2000). Specifically, they rated their agreement with the following statements: “Sometimes, you think that technology systems are not designed for use by ordinary people,” “There should be caution in replacing important people-tasks with technology because new technology can breakdown or get disconnected,” “Many new technologies have health or safety risks that are not discovered until after people have used them,” “Technology always seems to fail at the worst possible time,” and “New technology makes it too easy for governments and companies to spy on people” on five-point scales (1 = “Strongly disagree,” and 5 = “Agree strongly”); statements presented in random order). We averaged these ratings into an index measuring discomfort with technology (higher numbers indicate greater discomfort; $\alpha = .70$).

Then, in a two-cell between-subjects design, participants read a real news article describing the failure of either an algorithm or a person in the calculation of social security benefits.

Poorly trained Universal Credit [algorithm/person] forces people into hunger and debt. A new report found that the government benefit system—which is reliant on [an algorithm that/a person who] calculates benefits for people on a low income or out of work—threatens the rights of people most at risk of poverty.

People in the UK are being pushed into poverty by [a computer algorithm that/a person who] incorrectly allocated how much money they are entitled to in social security payments. Universal Credit claimants are being forced to forego food and take on debt because of a poorly trained algorithm, a leading human rights charity warned. A new report by Human Rights Watch found that the government benefit system—which is reliant on [an algorithm/a person] to calculate benefits for people on a low income or out of work—“threatens the rights of people most at risk of poverty”. The charity is calling on the government to take action ahead of an anticipated winter jobs crisis, with up to nine million workers at risk of redundancy as the furlough scheme is wound down.

Participants made an inferential judgment about the likelihood that [an algorithm/a person] employed by the state of Arkansas would make wrong social security benefits calculations (1 = “Unlikely to make wrong calculations,” and 7 = “Likely to make wrong calculations”).

Results and Discussion

We regressed inferential judgment on agent (dummy coded: AI = 1, human = 0), discomfort with technology ($M = 3.41$, $SD = .72$), and the interaction between agent and discomfort with technology (mean-centered). The analysis revealed a significant main effect of agent ($\beta = .144$, $t(395) = 2.96$, $p = .003$) and a significant two-way interaction between agent and discomfort with technology ($\beta = .181$, $t(395) = 2.68$, $p = .008$). There was no significant effect of discomfort with technology ($\beta = .050$, $t(395) = .74$, $p = .46$). Because discomfort with technology was a continuous measured moderator, we explored the interaction further using the Johnson–Neyman technique. The results revealed a positive and significant effect of agent on inferential judgment for levels of discomfort greater than 3.17 ($B_{JN} = .30$, $SE = .15$, $p = .050$). Figure 2 plots these results.

These results provide correlational evidence of the scope of algorithmic transference; the more participants reported discomfort with new technologies, the more they exhibited transference. The next study provided additional evidence of the scope of transference by exploring the role of human oversight.

Study 5: Human Oversight Eliminates Algorithmic Transference

Study 5 further delineated the scope of algorithmic transference and tested moderation by human oversight. Specifically, in this study we explored the case where AI is leveraged to assist and augment human intelligence, with a person acting in a role of oversight rather than AI fully replacing a person. We predicted that consumers would be less prone to transfer algorithmic failures when making an inference for an algorithm if this algorithm operates under human oversight. Indeed, having a human in the loop should prevent consumers from viewing the combination of an AI system and a human as an out-group, and thus eliminate transference.

Procedure

Respondents from MTurk participated in exchange for \$.35 ($N = 304$; $M_{\text{age}} = 41.6$ years, $SD = 12.6$; 52.3% female, 46.7% male, 1.0% nonbinary/third gender).

Participants read information about disability benefits allocated by states across the United States:

Residents living in the state of Michigan may sign up for the state disability program to help with their expenses. To assess eligibility for disability benefits, the state considers, among others, the following: the claimant’s ability to resume work or find work, the claimant’s income situation, the severity of the claimant’s impairment, the claimant’s medical condition, and the claimant’s ability to perform past work or any type of work. On average, between 35 and 40 percent of disability benefits calculations are inaccurate. States—which are charged with distributing and overseeing many federally funded benefits—are taking these benefits calculations seriously.

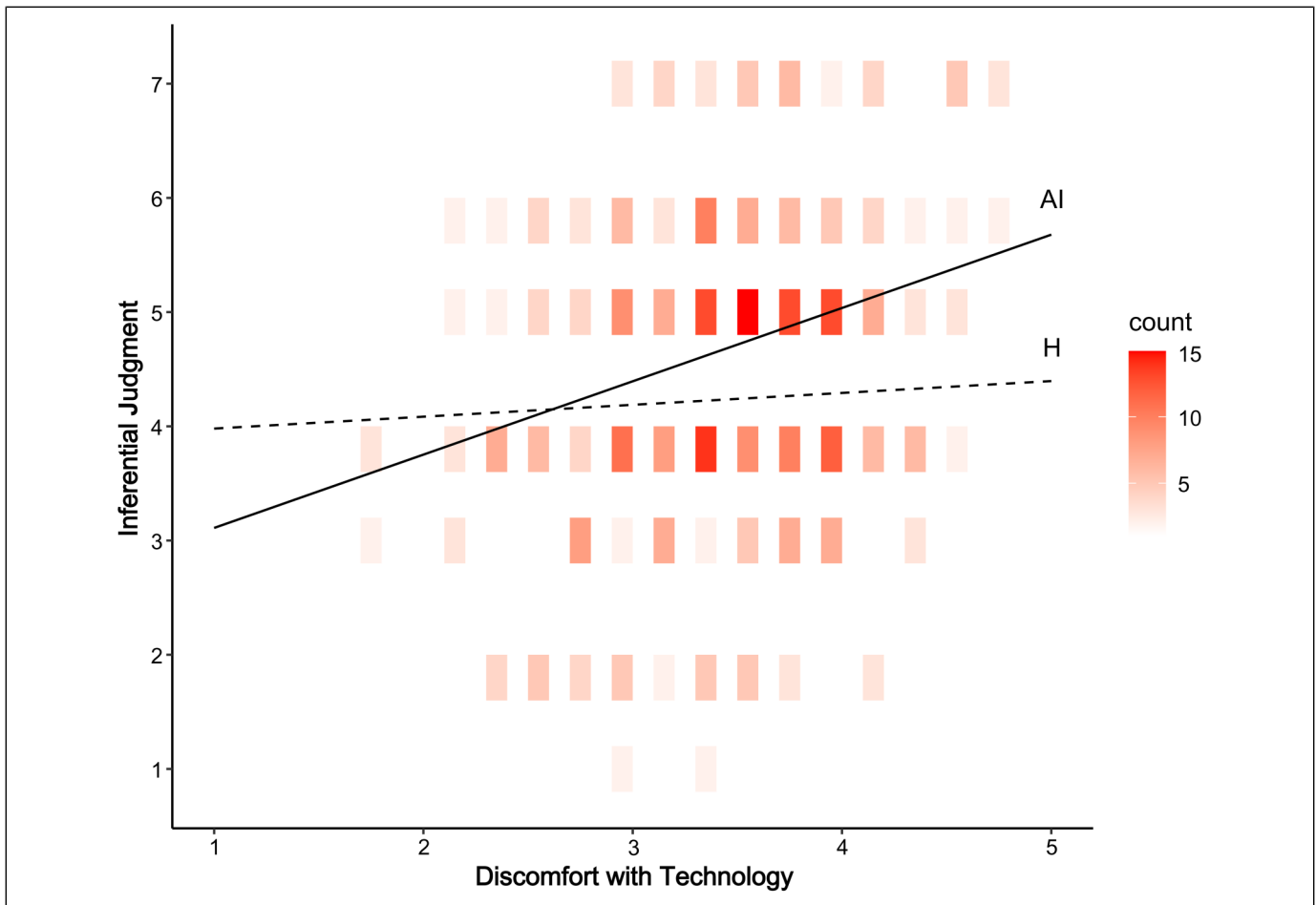


Figure 2. Results of Study 4: Greater Discomfort with New Technologies is Associated with Greater Algorithmic Transference.

Notes: Boxes represent collections of data points; the color saturation of the box indicates the number of data points in the range covered by the box. Regression lines represent the effect of agent on inferential judgment (plotted on the y-axis) as a function of discomfort with technology (plotted on the x-axis). There is a positive and significant effect of agent on inferential judgment for levels of discomfort greater than 3.17 ($\beta_N = .30$, $SE = .15$, $p = .050$).

Participants were randomly assigned to one condition in a three-cell between-subjects design. That is, participants read about a failure due to an algorithm, and then made an inferential judgment about another algorithm (algorithm condition); read about a failure due to a person, and then made an inferential judgment about another person (human condition); or read about a failure due to an algorithm, and then made an inferential judgment about another algorithm operating under human oversight (human oversight condition). Specifically, participants in the algorithm condition read:

In the recent past, the state of Michigan employed an algorithm to calculate disability benefits. Given the algorithm's calculations of the amount and type of disability benefits, the state allocated disability benefits to its residents. A state review later determined that most of these calculations—completed by the Michigan algorithm—were wrong. Now imagine that the state of Ohio were to employ an algorithm to calculate disability benefits. The algorithm would make a final determination about the disability benefits to allocate. This way

of deciding is technically called “artificial intelligence,” because it uses an algorithm to replace what human intelligence can do.

Participants in the human condition read:

In the recent past, the state of Michigan employed a person to calculate disability benefits. Given the person's calculations of the amount and type of disability benefits, the state allocated disability benefits to its residents. A state review later determined that most of these calculations—completed by the Michigan person—were wrong. Now imagine that the state of Ohio were to employ a person to calculate disability benefits. The person would make a final determination about the disability benefits to allocate. This way of deciding is technically called “human intelligence,” because it uses what human intelligence can do.

Participants in the human oversight condition read:

In the recent past, the state of Michigan employed an algorithm to calculate disability benefits. Given the algorithm's calculations of

the amount and type of disability benefits, the state allocated disability benefits to its residents. A state review later determined that most of these calculations—completed by the Michigan algorithm—were wrong. Now imagine that the state of Ohio were to employ an algorithm to calculate disability benefits. The algorithm would make an initial calculation and assist a person who would make the final determination about the disability benefits to allocate. This way of deciding is technically called “augmented intelligence,” because it uses algorithms to enhance and augment what human intelligence can do.

Participants made an inferential judgment about the likelihood of an incorrect decision in calculating disability benefits by the agent described in the condition: an algorithm, a person, or an algorithm under a person’s oversight (1 = “Unlikely to make wrong calculations,” and 7 = “Likely to make wrong calculations”).

Results and Discussion

A one-way analysis of variance on inferential judgment was significant ($F(2, 301) = 17.99, p < .001, \eta_p^2 = .11$). Planned contrasts between the algorithm and human conditions replicated algorithmic transference: participants transferred an algorithmic failure to a greater extent than a human failure ($M_{AI} = 5.39, SD = 1.37; M_H = 4.60, SD = 1.31; t(301) = 4.18, p < .001, d = .59$). However, when participants made an inferential judgment about an algorithm operating under human oversight, algorithmic transference was eliminated: participants made the same inferential judgment as they did when the failing agent was human ($M_{human\ oversight} = 4.30, SD = 1.36; t(301) = 1.56, p = .119, d = .22$), and a lower inferential judgment than they did when the failing agent was an algorithm ($t(301) = 5.77, p < .001, d = .81$).

These results delineated a boundary condition for algorithmic transference: having a person act in an oversight role eliminated failure transference from one algorithm to another one.

Studies 6a–6c: Implications for Propensity to Apply for Public Services

Studies 6a–6c tested the downstream consequences of algorithmic transference on consumer propensity to apply for public services. We assessed propensity to apply for public services as a proxy for behavioral consequences for consumers (Cian, Longoni, and Krishna 2020), as the provision of these services is the statutory duty of public agencies, and governments are grappling with underutilization of these benefits (Zipperer and Gould 2020).

Study 6a: Propensity to Apply for Disability Benefits

Procedure. Respondents from a nationally representative U.S. sample recruited via Prolific participated in exchange for \$.35 ($N = 299; M_{age} = 44.4$ years, $SD = 16.6$; 49.2% female, 48.1% male, 2.7% nonbinary/third gender).

Participants read information about the Social Security Disability program provided by the U.S. government taken from the government website Benefits.gov. Then, in a two-cell, between-subjects design, participants read about the failure of either an algorithm or a person employed by the state of Michigan in the calculation of these benefits:

Residents living in the state of Michigan may sign up for the state Social Security Disability program to help with their expenses. In the recent past, the state of Michigan employed [an algorithm/a person] to calculate disability benefits. The state then allocated Social Security disability benefits to its residents based on [the algorithm’s/the person’s] calculations. As it turns out, a state review later determined that most of the benefits calculations completed by the Michigan [algorithm/person] were wrong.

Participants then made an inferential judgment about the likelihood that [an algorithm/a person] employed by their state would make wrong benefits calculations (1 = “Unlikely to make wrong calculations,” and 7 = “Likely to make wrong calculations”) and rated their propensity to apply for Social Security Disability program benefits if the state where they lived employed an algorithm/person and they needed these benefits (1 = “Likely to apply,” and 7 = “Unlikely to apply”).

Results and discussion. A t-test on inferential judgment showed that participants were more prone to transfer an algorithmic failure than a human failure, replicating algorithmic transference ($M_{AI} = 5.32, SD = 1.59; M_H = 4.77, SD = 1.52; t(297) = 3.01, p = .003, d = .35$). Learning about an algorithmic failure was also associated with lower propensity to apply for disability benefits in case of need ($M_{AI} = 3.15, SD = 2.12; M_H = 2.63, SD = 1.80; t(297) = 2.32, p = .021, d = .27$). A mediation analysis (PROCESS Model 4; Hayes 2018) tested whether the relationship between agent and propensity to apply for benefits was mediated by inferential judgment. The mediational path from agent ($AI = 1$ and $human = 0$) to intention to apply was indeed significant through inferential judgment (indirect effect: .13, 95% CI: [.02, .27]; direct effect: .40, 95% CI: [−.05, .85]; total effect: .52, 95% CI: [.08, .98]).

Study 6b: Propensity to Apply for TANF Benefits

Procedure. The Temporary Assistance for Needy Families (TANF) benefit program is a state-administered program that provides families with financial assistance. To maximize relevance of the domain employed (temporary income support for families comprising at least one minor child), we relied on Cloud Research to recruit respondents who had at least one child below the age of 18 years or were pregnant. They participated in exchange for \$.35 ($N = 201, M_{age} = 40.6$ years, $SD = 8.3$; 56.2% female, 43.3% male, .5% nonbinary/third gender).

Participants read information about TANF benefits available on the U.S. government website Benefits.gov. Then, in a two-cell, between-subjects design, participants read about the

failure of either an algorithm or a person employed by the state of Michigan in the calculation of TANF benefits.

In the recent past, the state of Michigan employed [an algorithm/a person] to calculate TANF benefits to its residents. The state then allocated TANF benefits to its residents based on [the algorithm's/the person's] calculations. As it turns out, state review later determined that most of these TANF calculation decisions completed by the Michigan [algorithm/person] were wrong.

Participants then made an inferential judgment about the likelihood that [an algorithm/a person] employed by the state where they lived would make wrong TANF benefits calculations (1 = "Unlikely to make wrong calculations," and 7 = "Likely to make wrong calculations") and rated their propensity to apply for TANF benefits if the state where they lived employed [an algorithm/a person] and they needed these benefits (1 = "Likely to apply," and 7 = "Unlikely to apply").

Results and discussion. A t-test on inferential judgment replicated algorithmic transference ($M_{AI} = 5.24$, $SD = 1.53$; $M_H = 4.24$, $SD = 1.63$; $t(199) = 4.47$, $p < .001$, $d = .63$). Learning about an algorithmic failure was also associated with lower propensity to apply ($M_{AI} = 3.97$, $SD = 1.99$; $M_H = 2.93$, $SD = 1.79$; $t(199) = 3.89$, $p < .001$, $d = .55$). A mediation analysis (PROCESS Model 4; Hayes 2018) tested whether the relationship between agent and propensity to apply was mediated by inferential judgment. The mediational path from agent ($AI = 1$ and $human = 0$) to propensity to apply was significant through inferential judgment (indirect effect: .31, 95% CI: [.09, .61]; direct effect: .74, 95% CI: [.20, 1.27]; total effect: 1.04, 95% CI: [.51, 1.57]).

Study 6c: Propensity to Apply for Consumer Protection Services

Procedure. Respondents from MTurk participated in exchange for \$.35 ($N = 200$, $M_{age} = 40.5$ years, $SD = 12.4$; 41.2% female, 58.3% male, .5% nonbinary/third gender). As this study focused on the Consumer Financial Protection Bureau (CFPB), a bureau within the Federal Reserve System that protects consumers in the financial marketplace, we asked participants to read the following information:

The Bureau of Consumer Financial Protection, known as the Consumer Financial Protection Bureau (CFPB), is an independent bureau within the Federal Reserve System that makes sure banks, lenders, and other financial companies treat consumers fairly. The CFPB was created to provide a single point of accountability for enforcing federal consumer financial laws and protecting consumers in the financial marketplace. Please review the information below from the CFPB's website.

Participants also reviewed information about these services as per the description provided in the CFPB website. Then, in a two-cell, between-subjects design, participants read about

the failure of either a natural language processing algorithm or a person employed by the CFPB. Specifically, they read:

Imagine you read in the news that in the recent past, the Consumer Financial Protection Bureau has employed [a natural language processing algorithm/a person] to categorize narratives, identify trends, and predict consumer harm in textual consumer complaints. The [algorithm/person] handled consumer complaints about credit cards, mortgages, student loans, bank accounts or services, vehicle and consumer loans, credit reporting, debt collection, and money transfers. An audit by the Federal Reserve System later revealed that the [algorithm/person] erroneously handled most of these consumer complaints.

Participants then read that the Consumer Protection Agency of the state in which they lived is another institution that enforces laws to protect consumers from fraud, deception, and other unfair business practices. Participants rated the likelihood that [a natural language processing algorithm/a person] employed by the Consumer Protection Agency of their state would erroneously handle consumer complaints (1 = "Unlikely to erroneously handle consumer complaints," and 7 = "Likely to erroneously handle consumer complaints") and rated their propensity to apply for consumer protection services from the Consumer Protection Agency of their state if it employed [a natural language processing algorithm/a person] to handle claims (1 = "Definitively submit," and 7 = "Definitively not submit").

Results and discussion. A t-test on inferential judgment replicated algorithmic transference ($M_{AI} = 5.32$, $SD = 1.44$; $M_H = 4.40$, $SD = 1.43$; $t(198) = 4.57$, $p < .001$, $d = .65$). A t-test on propensity to apply also showed that learning about an algorithmic failure led to lower propensity to apply for consumer protection services than learning about a human failure ($M_{AI} = 4.36$, $SD = 1.56$; $M_H = 3.56$, $SD = 1.51$; $t(198) = 3.68$, $p < .001$, $d = .52$). A mediation analysis (PROCESS Model 4; Hayes 2018) tested whether the relationship between agent and propensity to apply was mediated by inferential judgment. The mediational path from agent ($AI = 1$ and $human = 0$) to propensity to apply for consumer protection services was indeed significant through inferential judgment (indirect effect: .19, 95% CI: [.04, .38]; direct effect: .61, 95% CI: [.16, 1.05]; total effect = .80, 95% CI: [.37, 1.23]).

Overall, Studies 6a–6c replicated algorithmic transference. More importantly, these studies highlighted that algorithmic transference may harm consumers and undermine their propensity to utilize public services.

General Discussion

Across a range of policy areas and diverse samples, we explored how people respond to failures of AI in the government. A series of preregistered studies documented a robust effect of algorithmic transference: higher generalization of algorithmic than human failures. In Studies 1a–1c, participants transferred algorithmic failures

at a higher rate than human failures. Study 2 directly tested group homogeneity perception as process evidence via mediation along with possible alternative explanations (general knowledge of AI and algorithms, perceived locus of causality). Study 3 tested our process account through moderation, showing that algorithmic transference is eliminated when out-group heterogeneity is salient. Delineating the scope of the phenomenon, Study 4 indicated that algorithmic transference is accentuated for consumers who are more uncomfortable with new technologies. Study 5 tested a case in which a human retains a role of oversight, which eliminated algorithmic transference. Finally, Studies 6a–6c tested the implications of algorithmic transference for propensity to utilize public services. Table 1 summarizes the empirical package, details of empirical testing, and results.

Contributions to Theory

Our research makes theoretical contributions to the literatures on psychological responses to automated systems, brand harms and brand scandals, and decision making in the public sector.

Contribution to research on responses to algorithmic failures and to AI systems. Since Paul Meehl's (1954) comparison between clinical and actuarial forecasting models, empirical investigations have explored responses to automated agents, delineating conditions under which people exhibit algorithm aversion (Cadario, Longoni, and Morewedge 2021; Dietvorst, Simmons, and Massey 2015; Longoni, Bonezzi, and Morewedge 2019, 2020; Longoni et al. 2022), appreciation (Castelo, Bos, and Lehmann 2019; Longoni and Cian 2020; Logg, Minson, and Moore 2019), or indifference (Berger et al. 2021). We extend this literature in the following ways.

First, our research makes a theoretical contribution to the literature on consumer responses to algorithmic failures. Research in this area has examined consumer responses on two dimensions. One dimension of consumer response has assessed reliance on an algorithmic versus a human advisor before and after learning of the advisor failure rate (Berger et al. 2021; Dietvorst, Simmons, and Massey 2015; Prah and Van Swol 2017; Srinivasan and Sarial-Abi 2021). Another dimension has assessed moral judgments of a failing algorithmic or human agent (Awad et al. 2020; Gill 2020). We contribute to this literature by extending the examination of consumer responses to inferential judgments and identifying the novel effect of algorithmic transference.

Our research makes a second theoretical contribution to the broader literature on consumer responses to AI systems (Cadario, Longoni, and Morewedge 2021; Castelo, Bos, and Lehmann 2019; Gill 2020; Longoni, Bonezzi, and Morewedge 2019, 2020). We show that consumers view AI systems as social agents and investigate the consequences of this social categorization process. Our focus is on processes stemming from how people assign social categories and mentally represent a group of artificial agents—as a group of a higher degree of homogeneity than a comparable group of humans. This perspective is novel, as it is based on group-level

representational processes due to social categorization (in-group vs. out-group) rather than on lay beliefs at the level of individual agents. That is, whereas prior research focused on lay beliefs about what an AI agent is presumed capable of doing, we focus on perceptions of AI agents at the group level—how a collection of artificial agents is mentally represented—and on the consequences of these representations. Overall, by empirically documenting how social categorization processes drive responses to algorithmic failures, and the consequences of these responses, we add an important contribution toward explaining people's perceptions of AI systems.

Contribution to research on brand harms. Another contribution of our research pertains to the literature on responses to brand harm. Research in this area shows that people respond less negatively to algorithmic than human-caused brand harm: brand devaluation is less pronounced following harm caused by an algorithm than by a person, an effect due to lower attribution of the failure to an algorithm than to a person (Srinivasan and Sarial-Abi 2021). Our research departs from these findings on two dimensions. First, the focus of our investigation is on transference—on inferential judgments about a different algorithm/person rather than on judgments about the same erring algorithm/person. Second, the driver of these inferential judgments is based on social categorization processes rather than on differential responsibility of the failure. Indeed, and as shown in Study 2, algorithmic transference is not due to differential locus of causality, but rather driven by group homogeneity perceptions.

We also build on and extend the stream of research on the spillover of a scandal surrounding one brand to another brand. This research shows that a scandal spills over if the company involved in the scandal is typical (and thus more diagnostic) of the category and/or if the scandal pertains to an attribute that is typical (and thus more diagnostic) of the category. An example would be the case of a scandal related to Burger King's hamburger meat resulting in lower brand evaluations of Wendy's (Roehm and Tybout 2006). An interesting empirical question pertains to the understanding of the extent of transference between brands versus transference between algorithms or between humans (we thank an anonymous reviewer for this suggestion). The notion that failures are more diagnostic for algorithms than for humans is consistent with our theoretical model, which predicts algorithmic failures to be generalized more than human failures. A key point of departure, however, is that our theoretical model highlights the importance of whether a scandal is attributed to a decision made by an algorithm or a person. Our theoretical model would suggest that scandal transference would only occur if the scandal is attributed to a nonhuman (i.e., an algorithm) and not if the scandal is attributed to a human. To the extent that consumers believe that brand scandals are attributable to humans, there should be less transference between two brands than between two algorithms.

We tested this conjecture in an ancillary preregistered study (Study 7: $N = 299$; details in Web Appendix F) in which we compared transference of a scandal (erroneous calculation of unemployment benefits) from an algorithm to another

Table 1. Overview of the Study Designs and Results for the Main Dependent Variables.

Evidence of Algorithmic Transference					
Study 1a: Disability benefits (N = 205; MTurk)					
	AI	Human		Test Statistic	
Inferential judgment	5.32 (1.40)	4.84 (1.23)		$t(203) = 2.60, p = .01, d = .36$	
Study 1b: Social security benefits (N = 300; nationally representative U.S. sample, Prolific)					
	AI	Human		Test Statistic	
Inferential judgment	4.81 (1.52)	4.30 (1.47)		$t(298) = 3.01, p = .003, d = .35$	
Study 1c: Fraud in benefit claims (N = 202; experts, Cloud Research/MTurk)					
	AI	Human		Test Statistic	
Inferential judgment	5.30 (1.44)	4.78 (1.59)		$t(200) = 2.41, p = .017, d = .34$	
Evidence of Group Homogeneity Perceptions as Process					
Study 2: Testing perceptions of group homogeneity as process (Unemployment benefits: N = 502; MTurk)					
	AI	Human		Test Statistic	
Inferential judgment	5.44 (1.18)	4.77 (1.25)		$t(500) = 6.13, p < .001, d = .55$	
Group homogeneity	5.60 (1.05)	4.46 (1.31)		$t(500) = 10.71, p < .001, d = .96$	
Study 3: Making out-group heterogeneity salient eliminates transference (Unemployment benefits: N = 403; MTurk)					
	Control		Heterogeneity Salient		Test Statistic
	AI	Human	AI	Human	
Inferential judgment	5.28 (1.23)	4.39 (1.46)	4.64 (1.32)	4.45 (1.49)	$F(2, 399) = 6.30, p = .011, \eta_p^2 = .02$
Evidence of the Scope of Algorithmic Transference					
Study 4: Discomfort with technology moderates transference (Social security benefits: N = 400; MTurk)					
	AI	Human		Test Statistic	
Inferential judgment	4.69 (1.61)	4.23 (1.33)		$t(398) = 3.16, p = .002, d = .32$	
Study 5: Human oversight eliminates transference (Disability benefits: N = 304; MTurk)					
	AI	Human	Human Oversight	Test Statistic	
Inferential judgment	5.39 (1.37)	4.60 (1.31)	4.30 (1.36)	$F(2, 301) = 17.79, p < .001, \eta_p^2 = .11$	
Implications of Algorithmic Transference for Consumers					
Study 6a: Disability benefits (N = 299; nationally representative U.S. sample, Prolific)					
	AI	Human		Test Statistic	
Inferential judgment	5.32 (1.59)	4.77 (1.52)		$t(297) = 3.01, p = .003, d = .35$	
Propensity to apply	3.15 (2.12)	2.63 (1.80)		$t(297) = 2.32, p = .021, d = .27$	
Study 6b: TANF benefits (N = 201; parents of minors, Cloud Research / MTurk)					
	AI	Human		Test Statistic	
Inferential judgment	5.24 (1.53)	4.24 (1.63)		$t(199) = 4.47, p < .001, d = .63$	
Propensity to apply	3.97 (1.99)	2.93 (1.79)		$t(199) = 3.89, p < .001, d = .55$	
Study 6c: Consumer protection services (N = 200; MTurk)					
	AI	Human		Test Statistic	
Inferential judgment	5.32 (1.44)	4.40 (1.43)		$t(198) = 4.57, p < .001, d = .65$	
Propensity to apply	4.36 (1.56)	3.56 (1.51)		$t(198) = 3.68, p < .001, d = .52$	

Notes: Inferential judgment was measured on a seven-point scale (1 = "Unlikely," and 7 = "Likely"). Group homogeneity perceptions were measured on seven-point scales (1 = "Disagree strongly," and 7 = "Agree strongly"). Propensity to apply was measured on a seven-point scale (1 = "Likely," and 7 = "Unlikely"). Standard deviations in parentheses.

algorithm, or from a person to another person, or from a brand to another brand. Participants made two judgments: they rated the likelihood that another algorithm/person/brand would fail (1 = “Unlikely to make wrong calculations,” and 7 = “Likely to make wrong calculations”), and they rated the extent to which, in their opinion, the decisions regarding the unemployment calculations were made by algorithms or by people (1 = “Made by algorithms,” and 7 = “Made by people”). Consistent with our theoretical model, participants transferred the scandal to another algorithm ($M_{AI}=5.14$, $SD=1.37$) to a greater degree than they transferred the scandal to another person ($M_H=4.46$, $SD=1.20$; $t(296)=3.66$, $p<.001$, $d=.52$) and to another brand ($M_B=4.21$, $SD=1.34$; $t(296)=5.03$, $p<.001$, $d=.71$). There was no difference in transference between the person and the brand conditions ($t(296)=1.34$, $p=.18$, $d=.19$). Corroborating our predictions, participants attributed the scandal in the algorithm condition to algorithmic decisions rather than human decisions ($M_{AI}=1.26$, $SD=.93$), whereas the scandals in the human and brand conditions were attributed more to human decisions ($M_H=6.78$, $SD=.78$; human vs. algorithm: $t(296)=34.66$, $p<.001$, $d=4.91$; $M_B=5.97$, $SD=1.51$; brand vs. algorithm: $t(1, 296)=29.75$, $p<.001$, $d=4.20$). These results showed that consumers transfer algorithmic failures to a greater extent than they transfer brand failures, underscoring the importance of considering social categorization processes—and the associated representations between out-groups of AI systems and in-groups of humans—as drivers of transference effects.

Contribution to research on decision making in the public sector. A final contribution of our research pertains to the theoretical understanding of people’s perceptions of what has been labeled the “automated administrative state” (Calo and Citron 2021). Despite growing and often heated debates about how the civil society should regulate AI, research and commentary in this area have mostly focused on the efficiency and effectiveness gains that AI might offer to public agencies (De Sousa et al. 2019). Little attention has been paid to how consumers respond to the automation of public agencies, even though they derive their legitimacy from the provision of services to the public (Calo and Citron 2021). Our research fills this gap, thus exploring the broader societal consequences of AI.

Implications for Practitioners and Public Policy

In 2019, the White House hosted the Summit on Artificial Intelligence in Government to accelerate adoption of AI in the government, highlighting innovative efforts of agencies that had already adopted AI and urging further AI applications that would render the government more effective, efficient, and responsive. In 2020, an executive order of the federal government (Executive Office of the President 2020) called on federal agencies to rely on AI to deliver public services and foster public trust in this technology. Along the same lines, the National Security Commission on Artificial Intelligence (2021) urged the United States to act now to field AI systems

and invest substantially more resources in AI innovation. Overall, these and many other government initiatives are intended to accelerate and support the use of AI across governmental agencies to carry out their statutorily committed duties (Calo and Citron 2021). Indeed, despite concerns about the short- and long-term impact of AI, much of the public sector has been, or will soon be, reshaped by AI.

From a public policy standpoint, our research underscores the importance of integrating the perspective of consumers in the assessment of the social impact of AI. Our results highlight how the hasty and mismanaged deployment of faulty AI systems may harm consumer uptake of public services. Consumers do not simply ascribe failure to the one governmental agency employing a faulty AI system. Instead, and perhaps unwarrantedly, consumers draw negative inferences about AI systems employed by other agencies. Thus, negative press about the failure of an AI system may have implications for other AI systems. Ultimately, and as shown in Studies 6a–6c, these negative inferences dampen consumer propensity to access public services, the provision of which is the statutory duty of public agencies.

Fortunately, our research also points to interventions that can mitigate this effect and boundary conditions under which the effect does not manifest. The results of Study 3 showed that it is possible to mitigate the effects of learning of failing algorithms by dispelling the belief that all AI systems are the same and by communicating that these systems are in fact heterogeneous. The results of Study 5 also showed that emphasizing having human oversight on algorithmic decision making prevents transference from emerging.

Overall, our research points to a novel and consequential implication of the reporting of failures of AI, at a time when actions are taken to enforce the transparent disclosure of AI systems executing government policies (e.g., Executive Order 13960 [Executive Office of the President 2020]). That is, government agencies are required to disclose proposed and existing AI systems, including their purpose, reach, use, and potential impact on communities and individuals. Disclosing the use of AI is intended to help governmental agencies proactively avoid backlash against systems the public may find untrustworthy.

Our research warns against the hurried and unregulated deployment of AI, as algorithmic transference may harm trust in the government, as evidence by an ancillary preregistered study that tested the implications of algorithmic transference for trust in the government (Study 8: $N=200$; details in Web Appendix G). Participants read about the failure of either an algorithm or a person employed by Allegheny County’s police department to monitor and prevent child abuse. Then, participants made two judgments. To measure transference, participants rated the likelihood that another [algorithm/person] employed by Rhode Island’s police department would fail (1 = “Unlikely to make wrong calculations,” and 7 = “Likely to make wrong calculations”). To measure trust in the government, participants rated the extent to which they would trust the Rhode Island police department if it employed [an algorithm/a person] (1 =

“Not at all,” and 7 = “Very much”). Replicating previous results, participants were more prone to transfer algorithmic failures than human failures ($M_{AI}=5.42$, $SD=.90$; $M_H=4.45$, $SD=1.05$; $t(198)=7.02$, $p<.001$, $d=.99$). An algorithmic failure was also associated with lower trust in the agency than the human failure was ($M_{AI}=2.52$, $SD=1.47$; $M_H=4.02$, $SD=1.17$; $t(198)=7.97$, $p<.001$, $d=1.13$). A mediation analysis showed that transference mediated the effect of agent on trust in the agency (indirect effect .36, 95% CI: [.11, .63]).

Limitations and Direction for Future Research

Despite the robustness of the phenomenon documented, our research has limitations that offer several opportunities for future research. We focus our discussion on five directions that we believe offer particularly promising directions to extend the current work.

First, it would be interesting to explore the antecedents of out-group homogeneity. Future research could examine why out-group homogeneity perceptions arise and how they operate, investigating, for instance, whether out-group homogeneity perceptions are the result of a cognitive bias or an error in social perception. Group perception is a flexible and dynamic process that may vary depending on the situational and social context (Palla, Barabasi, and Vicsek 2007). Mental representation of agents and their assignment to a group is indeed an adaptive feature; therefore, shifts in goals or strategies may render it beneficial to modify group member perceptions. Future research could track changes in how automated agents are mentally represented over time, and whether these changes accentuate or reduce out-group homogeneity and therefore transference.

Second, future research could dig deeper into the correlates of out-group homogeneity perceptions. We examined two such candidates in Studies 2 (algorithmic literacy) and 5 (discomfort with new technologies) and call for future research to examine other factors. For instance, research on intergroup perceptions links out-group homogeneity to need for uniqueness (Quattrone and Jones 1980), need for predictability (Irwin, Tripodi, and Bieri 1967), and need to justify in-group favoritism and out-group hostility (Wilder 1986). Future research could investigate whether perceptions of AI systems as homogeneous correlate with these needs.

Third, future research could further examine the role of knowledge of AI in moderating transference. Although the results of Study 1c (which employed a sample of computer/high-tech experts) and Study 2 are unresponsive of the notion that knowledge of AI drives transference, we note that we do not claim that consumers have as much knowledge of AI algorithms as they do of humans. As humans, we have more insight into the degree of variability that exists among humans, and less insight into what makes nonhuman agents different from each other. Furthermore, the algorithmic literacy scale we employed does not tap into everyday knowledge of AI, that is, consumer knowledge of digital technologies that employ AI systems such as Alexa and Siri. Our claim and what our studies provide evidence for is the hypothesis that group homogeneity is a unique driver of

algorithmic transference, even when accounting for knowledge of AI. We call for future research to dig deeper into what dimensions and types of AI knowledge could moderate transference.

Fourth, although we tested algorithmic transference across several policy areas, we focused on cases where there was an “accurate” value and decision correctness was unambiguous (i.e., fraud accusations are baseless). Thus, what would happen in the case of more ambiguous errors is an open question. Future research could explore cases in which it is unclear the extent to which a decision outcome is erroneous, and whether ambiguity in correctness moderates transference.

Finally, we focused on the government given the rapid spread of AI and its relevance for consumers given news media reports of AI failures. Nevertheless, algorithmic transference and its explanation likely apply to other consumer domains, such as recommendations from virtual assistants, customer service chatbots, and other automated company–customer interactions. In all these cases, algorithmic transference may harm companies other than the one employing a faulty AI system. For instance, Best Buy automated the process used to return products and hired The Retail Equation, a company that used an algorithm to score customers’ shopping behavior and impose limits on the amount of merchandise customers could return. The algorithm wrongfully banned many customers from returning products, a failure that ended up alienating Best Buy customers and, ultimately, resulted in a class-action lawsuit against not just Best Buy but also other companies that automated their refund policies (Bradley-Smith 2021). As this example suggests, future research could systematically investigate to what extent transference carries over to other marketing settings and consumption domains.

AI holds great potential to improve the public sector. However, we highlight how the deployment of AI in the government should integrate the consumer perspective and reflect the extent to which these technologies promote, rather than harm, the duty of public agencies toward consumers.

Acknowledgments

The authors would like to acknowledge the *JMR* review team for their guidance throughout the review process. The authors are grateful to seminar participants at the Max Planck Institute for Human Development, Cornell University, George Washington University, Wilfrid Laurier University, Bocconi University, the Association for Consumer Research, the Society for Consumer Psychology, and the Association for Computing Machinery Conference on Artificial Intelligence, Ethics, and Society for their engagement with this research and comments on earlier versions of this manuscript. The authors are indebted to Asifur Rahman, Becky Duff, and the Batten Institute for Entrepreneurship and Innovation for their research support. The authors gratefully acknowledge Dokyun Lee and Andrea Pellegrini for their help with the algorithmic literacy scale and Bruno Radice and Raina Zhang for their research assistance.

Associate Editor

Gergana Nenkov


Declaration of Conflicting Interests


The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.


Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iDs

Chiara Longoni  <https://orcid.org/0000-0002-4945-4957>

Luca Cian  <https://orcid.org/0000-0002-8051-1366>

Ellie J. Kyung  <https://orcid.org/0000-0003-2385-9523>

References

- Administrative Conference of the United States (2020), "Government by Algorithm: Artificial Intelligence in Federal Administrative Agencies," (February 19), <https://www.acus.gov/report/government-algorithm-artificial-intelligence-federal-administrative-agencies>.
- AIAAIC (2022), "AIAAIC Repository," (accessed May 30, 2022), <https://www.aiaaic.org/aiaaic-repository>.
- Awad, Edmond, Sydney Levine, Max Kleiman-Weiner, Sohan Dsouza, Joshua B. Tenenbaum, Azim Shariff, et al. (2020), "Drivers Are Blamed More Than Their Automated Cars When Both Make Mistakes," *Nature Human Behaviour*, 4 (2), 134–43.
- Aydinoğlu, Nilüfer and Luca Cian (2014), "Show Me the Product, Show Me the Model: Effect of Picture Type on Attitudes Toward Advertising," *Journal of Consumer Psychology*, 24 (4), 506–19.
- Berger, Benedikt, Martin Adam, Alexander Rühr, and Alexander Benlian (2021), "Watch Me Improve—Algorithm Aversion and Demonstrating the Ability to Learn," *Business & Information Systems Engineering*, 63 (1), 55–68.
- Bigman, Yochanan E., Kai C. Yam, Débora Marciano, Scott J. Reynolds, and Kurt Gray (2021), "Threat of Racial and Economic Inequality Increases Preference for Algorithm Decision-Making," *Computers in Human Behavior*, 122 (8), 106859.
- Borau, Sylvie, Tobias Otterbring, Sandra Laporte, and Samuel Fosso Wamba (2021), "The Most Human Bot: Female Gendering Increases Humanness Perceptions of Bots and Acceptance of AI," *Psychology & Marketing*, 38 (7), 1052–68.
- Bradley-Smith, Anna (2021), "Best Buy, Dick's Sporting Goods, Other Major Chain's Software and Analytics Service Provider Invades Consumers' Privacy and Falsely Labels Them Fraudsters, Class Action Alleges," Top Class Actions (August 2), <https://topclassactions.com/lawsuit-settlements/privacy/best-buy-dicks-sporting-goods-other-major-chains-software-and-analytics-service-provider-invades-consumers-privacy-and-falsely-labels-them-fraudsters-class-action-alleges/>.
- Brewer, Marylinn B. (1993), "Social Identity, Distinctiveness, and In-Group Homogeneity," *Social Cognition*, 11 (1), 150–64.
- Cadario, Romain, Chiara Longoni, and Carey K. Morewedge (2021), "Understanding, Explaining, and Utilizing Medical Artificial Intelligence," *Nature Human Behavior*, 5 (12), 1636–42.
- Calo, Ryan and Danielle K. Citron (2021), "The Automated Administrative State: A Crisis of Legitimacy," *Emory Law Journal*, 70 (4), 797–845.
- Castelo, Noah, Maarten W. Bos, and Donald R. Lehmann (2019), "Task-Dependent Algorithm Aversion," *Journal of Marketing Research*, 56 (5), 809–25.
- Cian, Luca, Chiara Longoni, and Aradhna Krishna (2020), "Advertising a Desired Change: When Process Simulation Fosters (vs. Hinders) Credibility and Persuasion," *Journal of Marketing Research*, 57 (3), 489–508.
- De La Garza, Alejandro (2020), "States' Automated Systems Are Trapping Citizens in Bureaucratic Nightmares with Their Lives on the Line," *Time* (May 28), <https://time.com/5840609/algorithm-unemployment/>.
- De Sousa, Wesley G., Elis Regina Pereira de Melo, Paulo H. De Souza Bermejo, Rafael A. Sousa Farias, and Adalmir O. Gomes (2019), "How and Where Is Artificial Intelligence in the Public Sector Going? A Literature Review and Research Agenda," *Government Information Quarterly*, 36 (4), 101392.
- Dietvorst, Berkeley J., Joe P. Simmons, and Cady Massey (2015), "Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err," *Journal of Experimental Psychology: General*, 144 (1), 114–26.
- Dzindolet, Mary T., Linda Pierce, Hall P. Beck, and Lloyd Dawe (2002), "The Perceived Utility of Human and Automated Aids in a Visual Detection Task," *Journal of the Human Factors and Ergonomics Society*, 44 (1), 79–94.
- Executive Office of the President (2020), "Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government," Executive Order 13960 (December 3), 85 *Federal Register* 78939, <https://www.federalregister.gov/d/2020-27065>.
- Fogg, B.J.J. and Clifford Nass (1997), "How Users Reciprocate to Computers: An Experiment That Demonstrates Behavior Change," in *CHI '97 Extended Abstracts on Human Factors in Computing Systems*. New York: Association for Computing Machinery, 331–32.
- Gill, Tripat (2020), "Blame It on the Self-Driving Car: How Autonomous Vehicles Can Alter Consumer Morality," *Journal of Consumer Research*, 47 (2), 272–91.
- Goodwin, Henry (2020), "'Poorly Designed' Universal Credit Algorithm Forcing People into Hunger and Debt," *The London Economic* (September 29), <https://www.thelondoneconomic.com/politics/poorly-designed-universal-credit-algorithm-forcing-people-into-hunger-and-debt-203635/>.
- Granulo, Armin, Christoph Fuchs, and Stefano Puntoni (2020), "Preference for Human (vs. Robotic) Labor Is Stronger in Symbolic Consumption Contexts," *Journal of Consumer Psychology*, 31 (1), 72–80.
- Hayes, Andrew F. (2018), *Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-Based Approach*, 2nd ed. New York: Guilford Press.
- Hogg, Michael A. and Scott A. Reid (2006), "Social Identity, Self-Categorization, and the Communication of Group Norms," *Communication Theory*, 16 (1), 7–30.
- Irwin, Marc, Tony Tripodi, and James Bieri (1967), "Affective Stimulus Value and Cognitive Complexity," *Journal of Personality and Social Psychology*, 5 (4), 444–48.
- Lecher, Colin (2018), "What Happens When an Algorithm Cuts Your Health Care," *The Verge* (March 21), <https://www.theverge.com/2018/3/21/17144260/healthcare-medicaid-algorithm-arkansas-cerebral-palsy>.

- Linville, Patricia W. and Gregory W. Fischer (1998), "Group Variability and Covariation: Effects on Intergroup Judgment and Behavior," in *Intergroup Cognition and Intergroup Behavior*, C. Sedikides and Chester A. Insko, eds. Mahwah, NJ: Lawrence Erlbaum Associates, 123–50.
- Logg, Jennifer M., Julia Minson, and Dan A. Moore (2019), "Algorithm Appreciation: People Prefer Algorithmic to Human Judgment," *Organizational Behavior and Human Decision Processes*, 151, 90–103.
- Longoni, Chiara, Andrea Bonezzi, and Carey K. Morewedge (2019), "Resistance to Medical Artificial Intelligence," *Journal of Consumer Research*, 46 (4), 629–50.
- Longoni, Chiara, Andrea Bonezzi, and Carey K. Morewedge (2020), "Resistance to Medical Artificial Intelligence Is an Attribute in a Compensatory Decision Process: Response to Pezzo and Beckstead," *Judgment and Decision Making*, 15 (3), 446–48.
- Longoni, Chiara and Luca Cian (2020), "Artificial Intelligence in Utilitarian vs. Hedonic Contexts: The 'Word-of-Machine' Effect," *Journal of Marketing*, 86 (1), 91–108.
- Longoni, Chiara, Andrey Fradkin, Luca Cian, and Gordon Pennycook (2022), "News from Generative Artificial Intelligence Is Believed Less," in *FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency*. New York: Association for Computing Machinery, 97–106.
- Meehl, Paul, E. (1954), *Clinical Versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence*. Minneapolis: University of Minnesota Press.
- Nass, Clifford and Youngme Moon (2000), "Machines and Mindlessness: Social Response Computers," *Journal of Social Issues*, 56 (1), 81–103.
- Nass, Clifford, Youngme Moon, and Nancy Green (1997), "Are Machines Gender Neutral?" *Journal of Applied Social Psychology*, 27 (10), 864–76.
- Nass, Clifford, Jonathan Steuer, and Ellen R. Tauber (1994), "Computers Are Social Actors," in *CHI '94: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. New York: Association for Computing Machinery, 72–78.
- National Security Commission on Artificial Intelligence (2021), "Final Report," (accessed September 14, 2022), <https://www.nscai.gov/wp-content/uploads/2021/03/Full-Report-Digital-1.pdf>.
- Nisbett, Richard E., David H. Krantz, Christopher Jepson, and Ziva Kunda (1983), "The Use of Statistical Heuristics in Everyday Inductive Reasoning," *Psychological Review*, 90 (4), 339–63.
- Ostrom, Thomas M. and Constantine Sedikides (1992), "Outgroup Homogeneity Effects in Natural and Minimal Groups," *Psychological Bulletin*, 112 (3), 536–52.
- Palla, Gergely, Albert-László Barabasi, and Tamás Vicsek (2007), "Quantifying Social Group Evolution," *Nature*, 446 (7136), 664–67.
- Parasuraman, Ananthanarayanan (2000), "Technology Readiness Index (TRI) a Multiple-Item Scale to Measure Readiness to Embrace New Technologies," *Journal of Service Research*, 2 (4), 307–20.
- Park, Bernadette and Reid Hastie (1987), "Perception of Variability in Category Development: Instance- Versus Abstraction-Based Stereotypes," *Journal of Personality and Social Psychology*, 53 (4), 621–35.
- Park, Bernadette, Charles M. Judd, and Carey S. Ryan (1991), "Social Categorization and the Representation of Variability Information," *European Review of Social Psychology*, 2 (1), 211–45.
- Prahl, Andrew and Lyn Van Swol (2017), "Understanding Algorithm Aversion: When Is Advice from Automation Discounted?" *Journal of Forecasting*, 36 (6), 691–702.
- Quattrone, George A. and Edward E. Jones (1980), "The Perception of Variability Within Ingroups and Outgroups: Implications for the Law of Small Numbers," *Journal of Personality and Social Psychology*, 38 (1), 141–52.
- Reeves, Byron and Clifford Nass (1996), *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*. Cambridge, UK: Cambridge University Press.
- Roehm, Michelle L. and Alice M. Tybout (2006), "When Will a Brand Scandal Spill Over, and How Should Competitors Respond?" *Journal of Marketing Research*, 43 (3), 366–73.
- Rossen, Brent, Kyle J. Johnson, Adeline Deladisma, David S. Lind, and Benjamin Lok (2008), "Virtual Humans Elicit Skin-Tone Bias Consistent with Real-World Skin-Tone Biases," in *IVA '08: Proceedings of the 8th International Conference on Intelligent Virtual Agents*, Helmut Prendinger, James Lester, and Mitsuru Ishizuka, eds. Berlin: Springer.
- Ryan, Richard M. and James P. Connell (1989), "Perceived Locus of Causality and Internalization: Examining Reasons for Acting in Two Domains," *Journal of Personality and Social Psychology*, 57 (5), 749–61.
- Sedikides, Constantine and Thomas M. Ostrom (1993), "Perceptions of Group Variability: Moving from an Uncertain Crawl to a Purposeful Stride," *Social Cognition*, 11 (1), 165–74.
- Sonal, Pandya, Luca Cian, and Rajkumar Venkatesan (2022), "Grocery Shopping for America," *PsyArXiv* (September 13), psyarxiv.com/8r23e.
- Srinivasan, Raji and Gülen Sarial-Abi (2021), "When Algorithms Fail: Consumers' Responses to Brand Harm Crises Caused by Algorithm Errors," *Journal of Marketing*, 85 (5), 74–91.
- Tajfel, Henri (1969), "Cognitive Aspects of Prejudice," *Journal of Social Issues*, 25 (4), 79–97.
- Tay, Benedict, Younbo Jung, and Tazoon Park (2014), "When Stereotypes Meet Robots," *Computers in Human Behavior*, 38, 75–84.
- Turner, John C. (1982), "Towards a Cognitive Redefinition of the Social Group," in *Social Identity and Intergroup Relations*, H. Tajfel, ed. Cambridge, UK: Cambridge University Press, 15–36.
- Wilder, David A. (1986), "Social Categorization: Implications for Creation and Reduction of Intergroup Conflict," *Advances in Experimental Social Psychology*, 19, 291–355.
- Zipperer, Ben and Elise Gould (2020), "Unemployment Filing Failures: New Survey Confirms That Millions of Jobless Were Unable to File an Unemployment Insurance Claim," Economic Policy Institute (April 28), <https://www.epi.org/blog/unemployment-filing-failures-new-survey-confirms-that-millions-of-jobless-were-unable-to-file-an-unemployment-insurance-claim/>.