Press ‘1’ to speak to a machine: An examination of the psychological factors influencing preference for interaction with artificially intelligent actors

by

Hee Jin (Heather) Yang

S.M. Management Research
Massachusetts Institute of Technology, 2017

B.A. Psychology
Carleton College, 2012

SUBMITTED TO THE SLOAN SCHOOL OF MANAGEMENT IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY IN MANAGEMENT

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

JUNE 2021

Signature of Author:____
Department of Management
May 7, 2021

Certified by: ____
John Stephen Carroll
Gordon Kaufman Professor of Management, Emeritus
Professor Post-Tenure of Work and Organization Studies
Thesis Supervisor

Accepted by: ____
Catherine Tucker
Sloan Distinguished Professor of Management
Professor of Marketing
Chair, MIT Sloan PhD Program

Press '1' to speak to a machine: An examination of the psychological factors influencing preference for interaction with artificially intelligent actors

by

Hee Jin Yang

Submitted to the Sloan School of Management on May 7, 2021 in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy in Management

ABSTRACT

What psychological factors influence the preference for interaction with a human versus an artificially intelligent actor? How can these factors be used to increase adoption of novel technologies, and what are their broader societal impacts? In this dissertation, I answer these questions through two streams of research: Firstly, by examining what kinds of people seek out algorithmic advice; and secondly, how the implicit application of social information to algorithmic agents impacts their interpretability and evaluation.

In Chapter 1, I examine the individual level differences of users of artificially intelligent advisors. Across four studies, users' cognitive style predicted advice-seeking behavior from algorithmic advisors, even after controlling for a host of consequential factors, such as prior experience with artificial intelligence, comfort with technology, social anxiety, and educational background. Building on the Dual Process theory literature, I show that increased cognitive reflection is related to increased perceptions of accuracy for algorithmic (versus human) advisors, with accuracy perceptions mediating the relationship between cognitive style and advisor preference. I find that individuals who rely on their intuition perceive human advisors as being more accurate than algorithmic advisors, in comparison to their deliberative counterparts, and also rate algorithmic advisors as being less impartial.

In Chapter 2, I investigate how individuals apply social stereotypes to digital voiced assistants (DVAs) and how this facilitates understanding of novel personified devices. Through experimentally pairing participants with fake artificially intelligent voiced agents, I demonstrate that individuals implicitly apply social stereotypes to the agent in the same way as they do to humans. Consistent with traditional gender stereotypes and in contrast to current academic justifications reliant on the generalized preference for female voices, I find that individuals prefer female (versus male) voiced artificial intelligent agents when occupying roles that are female-typed, but not male-typed, demonstrating a stereotype congruence effect. I extend this finding to show how gender stereotype congruent features of a novel device facilitate understanding of its capabilities for inexperienced users.

Finally, I discuss the implications of this research for managers, policy makers, developers and users of artificially intelligent agents.

Thesis Supervisor: John Stephen Carroll
Title: Gordon Kaufman Professor of Management, Emeritus

ACKNOWLEDGEMENTS

This work would not have been possible without the support of a whole community. I first want to acknowledge the tremendous amount of guidance my dissertation committee has given me. I have learned so much from each of them, not only the craft of doing rigorous research, but what it means to be an academic. They have poured hours of thought, effort, and energy into improving my work and helping me grow as a researcher, each in their unique ways: My co-author and tireless cheerleader Renée Richardson Gosline has been generous in praise and support ever since our first meeting at the Behavioral Lab. I have been privileged enough to benefit from Jackson Lu's brainpower in developing my research ideas. Similarly, Jared Curhan has been meticulous and thoughtful in his feedback on my work and has been invested in my progress since day one of my entry into the program. And luckily for me, Basima Tewfik joined Sloan just in the nick of time to let me bask in her wisdom, which has carried me through the final stages of the program to great success.

Finally, I could not have made it through the program without John Carroll, whose judgment I trust completely and whose opinion I respect the most. I will forever be grateful for how John always encouraged me to pursue my ideas – no matter how wild and amorphous – with the greater aim of creating an impactful and meaningful research agenda. Meeting with John always left me with a renewed sense of direction, purpose, and confidence that I was capable of tackling the immense amount of work ahead. Selflessness, like John's, is rare; the sheer volume of pages of work that he has read (and corrected) rivals only the countless headaches that I have brought him. The best decision I made at Sloan was to ask John to be my advisor.

The Sloan community at large has kept me going. I have benefited from the brilliance and generosity of spirit from Erin Kelly, JoAnne Yates, Roberto Fernandez, Tom Malone, Wanda Orlikowski, Ray Reagans, Nate Wilmers, Christian Catalini, Anjali Sastry, and Hiram Samel. I also could not have succeeded in the program without the camaraderie and support of my colleagues from Sloan and Harvard (in alphabetical order, in this non-exhaustive list): Brittany Bond, Avi and Manuela Collis, Erik Duhaime, Carolyn Fu, Simon Friis, Leroy Gonsalves, Rebecca Grunberg, David Hagmann, Karen Huang, Jenn Logg, Hagay Volvovsky, Duanyi Yang, Helen Yap.

In the wider world, I am grateful for the hospitality afforded to me by Eric Hehman and his wonderful lab at McGill University. Eric, Sally, Eugene, and Neil reinvigorated me for the final stretch with their enthusiasm, kindness, and brilliant minds. Of course, none of this would be possible without the unconditional love from my family. Their unwavering support has carried me through the best and toughest of times, even despite the distance.

Last, but not least, I have to thank my amazing research assistants, Hannah Schiller and Leonardo Trigo, who have inspired me with their persistence and tenacity. Without them, my dissertation would have taken even longer to complete.

DEDICATION

I dedicate this to P. You are my world, my everything.

TABLE OF CONTENTS

Press '1' to speak to a machine: An examination of the psychological factors influencing preference for interaction with artificially intelligent

INTRODUCTION

What psychological factors influence the preference for interaction with a human versus an artificially intelligent actor? How can these factors be used to increase adoption of novel technologies, and what are their broader societal impacts? In this dissertation, I answer these questions through two streams of research: Firstly, by examining what kinds of people seek out algorithmic advice; and secondly, how the implicit application of social information to algorithmic agents impacts their use.

In Chapter 1, with Dr. Renée Richardson Gosline, I examine the individual level differences of users of artificially intelligent advisors. People increasingly interact with artificially intelligent (AI) agents in both their personal and professional lives, including chatbots,[1] robo-advisors,[2] and virtual assistants (such as Amazon's Alexa or Apple's Siri). The dawn of digitally-mediated experiences portends the rise of greater user empowerment, as individuals are afforded the power to seek advice from multiple sources and choose the degree to which algorithms play a role in making decisions about their health, finances, dating, and other domains. Yet, research about attitudes toward artificially intelligent algorithmic input demonstrates mixed reactions. A report by the Center for the Governance on AI at Oxford

[1] A chatbot is a piece of software that conducts a conversation via auditory or textual methods. Such programs are often designed to convincingly simulate how a human would behave as a conversational partner.

[2] Robo-advisors are a class of financial advisor that provide financial advice or investment management online, based on mathematical rules or algorithms, with moderate to minimal human intervention.

University found mixed reactions to AI, with 41% supporting the development of AI agents compared to a smaller 22% who opposed it (Zhang & Dafoe, 2019). A recent survey by Forrester (2019) found that 54% of U.S. consumers think that interactions with chatbots will negatively impact the quality of their lives. On the positive side, evidence points to a growing desire for, and adoption of, technologically-mediated experiences: 70% of respondents said that they sought the power to solve their own customer service issues without having to talk to another person and would prefer to do so via text, chat, or messaging if "it were done right" (Aspect Customer Experience Survey, 2016). Moreover, 69% of these users said that they interact with an intelligent assistant or chatbot at least once a month.

Algorithmic advice is widespread and effectively employed across a variety of disparate decision domains with its usage predicted to only grow (Brynjolfsson & McAfee, 2011). In mortgage lending decisions, algorithmic decision-making is lauded as the future, with 45% of a sample of over 2,000 FinTech firms offering online or app-based mortgage contracting by 2018 (Bartlett et al., 2019). In education, algorithms are used to assign grades for high school examinations (Simonite, 2020) and to determine placement into high schools (Cassano, 2019). The trend of using predictive models for higher education admissions is on the rise, with private and public universities in 25 states across the nation currently using online data or algorithms to generate scores for student suitability (MacMillan & Anderson, 2019).

This prevalence reflects the benefits that algorithmic advisors currently, and potentially could, bring. A study of an online lending platform found that predictions of default risk from a machine learning algorithm outperformed human estimates, benefiting both the lenders -- by guarding against risky loans -- and the borrowers -- by providing greater access to capital to underserved individuals (Fu et al., 2020). In traditional bank loan decisions, algorithmic decision frameworks reduce the time to screen applicants by 12-50%, while still maintaining bank profits and minimizing default risk (Metawa et al., 2017). A meta-analysis for employee hiring and admissions decisions found that mechanical selection of candidates using a simple algorithm was 50% more effective in predicting job performance than human experts (Kuncel et al., 2013). In medical diagnoses, artificial intelligence has been found to be more accurate than human judgment when analyzing ultrasounds for the detection of cancerous lesions (see Golden, 2017; Rodriguez-Ruiz et al., 2019; Liu et al., 2018), and meta-analyses of over fifty years of data has documented the efficacy of statistical models outperforming clinical judgements (Ægisdóttir et al., 2006). Even as rudimentary actuarial formulae, algorithmic advice has been established as consistently outperforming human experts across a wide variety of domains (Dawes, Faust, & Meehl, 1989). Overall, across multiple different decision-making areas, there is much evidence for the superiority of algorithmic advisors, as compared to their human equivalents.

Cognitive style is a robust individual difference that has been associated with a host of consequential judgment and decision-making biases (Campitelli and Labollita, 2010, Toplak, West, & Stanovich, 2011), as measured through the Cognitive Reflection Test (Frederick, 2005). As a lack of cognitive reflection has been theorized as a person's tendency to over-rely on their automatic –*albeit at times incorrect*– System 1 responses, a reflective respondent must suppress their gut reaction to arrive at the correct answer. We found cognitive reflection to be positively related to greater openness to advice from an algorithmic advisor. Even after controlling for age, education, prior experience with AI, comfort with technology, and social anxiety, this consistent effect predicted how much advice users sought out from an algorithmic (versus a human) advisor in a hypothetical scenario.

This line of work identifies a key individual level difference that organizations can leverage to help initiatives to encourage (or dissuade) advice adoption from human versus algorithmic sources, as well as suggesting a mechanism to help tailor initiatives to convince employees or users who may initially be wary of one form of advice over the other. For practitioners rolling out novel AI-agent based features, these results suggest an easier adoption trajectory for target audiences naturally higher in cognitive reflection (e.g., non-religious, socially liberal, more critically thinking, low testosterone individuals, Shenhav, Rand, & Green, 2012; Bahcekapili & Yilmaz, 2017; Deppe et al., 2015; Iyer, Koleva, Graham, Ditto, & Haidt, 2012; Nadler, Jiao, Peiran, & Johnson, 2017) and also identify potential stalwarts who may be wary of algorithmic encroachment. Next directions in this work will seek to manipulate the mechanism of perceived accuracy by which cognitive reflection relates to algorithmic preference, including an investigation into how resistant individuals of varying cognitive styles are when confronted with algorithmic advice of varying degrees of accuracy.

In Chapter 2, I draw upon the stereotype congruence literature to establish that, while individuals may explicitly justify their preference for a Digital Voiced Assistant (DVA) due to superficial characteristics of a female voice, their preferences are driven by stereotype congruence. Gender cues, as conveyed through a device’s voice, can either be congruent or incongruent with the stereotypes associated with the role that it is designed to serve, such as the female-typed role of assistant. I propose that congruent pairings serve to facilitate understanding of novel technologies by borrowing from categorical information of the stereotype, such that novice users better understand a device’s capabilities. In this chapter, I investigate the facilitatory nature of stereotype congruent features of digital voiced agents, with a special focus on the match between the device’s voice gender and the gender-typing of the device’s job role. I discuss future directions for research and practical implications for user-interface designers and policy makers for artificially intelligent personified technology.

Together, the two chapters demonstrate that users of artificially intelligent technologies apply some of the same heuristics and biases that are utilized in interpersonal interaction in the evaluation of these new technologies.

CHAPTER 1: COGNITIVE REFLECTION AND ALGORITHMIC AVERSION

Who resists algorithmic advice?

Cognitive style correlates with algorithmic aversion

Heather Yang and Renée Richardson Gosline

Abstract

As technology and artificially intelligent (AI) algorithms become increasingly prevalent in all aspects of life, individuals have more opportunity to rely on them as sources of advice when making consequential decisions. Recent research has documented both *algorithmic aversion* and *appreciation*, but little is known about which types of individuals are likely to resist or prefer algorithmic (relative to human) advice and why. In this paper, we present the first evidence that cognitive style affects the kind of advice people prefer. We present results from four studies (combined *N* = 2,450) showing that cognitive style consistently predicts the degree to which decision makers seek input from AI versus human advisors. Individuals who are cognitively reflective (versus intuitive) embrace greater input from algorithms in their decision-making, demonstrating *algorithmic appreciation*. We find that this effect is mediated by decision makers' perceptions of the expected accuracy and impartiality of the advisor: cognitively intuitive individuals believe human advisors are more accurate than do their cognitively deliberative peers, and where cognitively deliberative individuals expect algorithmic advisors to be more impartial human ones, as compared to their cognitively intuitive counterparts, thus resulting in differential preferences. Reliance on algorithms may therefore

depend not simply on actual accuracy, but rather whether people believe that a rational, deliberative process can indeed lead to accurate decisions.

Introduction

Imagine you are the manager of a popular e-commerce retailer's fulfillment center. At the beginning of each workday, you are responsible for making sure that there are sufficient warehouse employees to retrieve and pack items to be shipped. If you estimate the demand for the day too low, you run the risk of overworking your already tired employees. However, if you estimate too high, you run the risk of using up the budget prematurely, leaving you with poor coverage for peak times in the future. The entire workforce assembles in the morning to hear how many of them will be dismissed for the day -- a number that you decide with help from a new algorithm that predicts that day's orders. Despite the algorithm having been created by scientists much smarter than yourself, you believe that your knowledge and experience on the floor have more to contribute and end up ignoring the forecast altogether, to disastrous results. Although one can hope that learning can take place in the scenario described above (based on a true, confidential example from a well-known e-retailer, from personal communication), resistance to algorithmic advice is a persistent and pervasive issue that decision-making scholars and practitioners have grappled with to this day.

Given the widespread and ever-growing impact of AI in all aspects of our daily lives, it is imperative that we better understand how individuals decide to incorporate algorithmic input in their decisions. In this paper, we examine how a person's cognitive style – whether they tend to rely on intuitive heuristics or on analytical deliberation to make their decisions (Kahneman, 2011) – shapes their preference for algorithmic advice over human advice, a phenomenon referred to as *algorithmic appreciation* (Logg et al., 2019). Using data from four studies, we test the association between an individual's cognitive style and their propensity to seek advice from an algorithmic source and develop two main findings. First, building upon research using the Cognitive Reflection Test (Frederick, 2005), we find that a person's tendency to rely on their automatic, but often incorrect, responses is related to their preference for human (versus algorithmic) input when making decisions; conversely, those with a more analytic cognitive style, who engage in greater reflection to find correct answers, are more likely to embrace input from algorithms. Second, we find that expectations of accuracy of the human advisor mediate the relationship between cognitive style and algorithmic aversion: intuitive thinkers expect human (versus algorithmic) advisers to be more accurate than do deliberative thinkers. To a smaller extent, we also find that expectations of algorithmic impartiality mediates the effect, with cognitively deliberative individuals believing in the impartiality of algorithmic advisors more so than their intuitive counterparts. We discuss the implications of this research, emphasizing both vulnerability to algorithmic bias due to a predilection for AI advisors, on the one hand, and predisposition to ignore algorithmic advice and thus miss out on the ever-increasing benefits that technology affords, on the other.

Theoretical Background

Algorithmic Appreciation

Recent research has detailed instances in which people prefer advice from an algorithm to that from another human. In the paper that introduced the term "algorithmic appreciation," Logg and colleagues (2019) showed that individuals are more likely to take advice for forecasting and estimation tasks when it is framed as being from an algorithm, rather than from other people, controlling for advice quality. Therapeutic chatbots that provide mental health support have been shown to be helpful to people less inclined or able to seek human advice, with one study experimentally showing that participants engaged more with a counseling service when described as being algorithmically operated than when described as being staffed by a person (Lucas et al., 2017). In the arena of financial forecasting, individuals were more likely to prefer algorithmically-derived estimates when the algorithm was more likely to generate a perfect forecast, even if it was wrong most of the time (Dietvorst & Bharti, 2020). Taken together, this stream of research has established the phenomenon of *algorithmic appreciation*: the preference for advice from algorithmic sources.

Algorithmic Aversion

However, whereas "algorithmic appreciation" was introduced into the literature relatively recently, skepticism of algorithmic superiority or "algorithmic aversion" dates back decades earlier to Paul Meehl's (1954) work on actuarial models in predicting human behavior. Meehl found that simple statistical models outperformed experts in the field, yet the experts resisted accepting the results, and researchers at the time tried repeatedly to explain away or limit the findings. In more recent research, we see a preference for human sources in subjective tasks, even to the detriment of accuracy (Yeomans, et al., 2019). For example, in predicting how funny someone would consider a joke or whom someone would find attractive for a date, participants preferred taking the bad advice of another person to that of a high-quality algorithm. Other research has shown that, when asked to choose between a healthcare recommendation from a human physician or one from a computer program, participants preferred the human, ostensibly due to the ability to shift responsibility for a poor outcome to another person as opposed to an unblameable tool (Promberger & Baron, 2006). Similarly, within the healthcare domain, Longoni, Bonezzi, and Morewedge (2019) identified "uniqueness neglect" in medical care patients who felt that, although algorithmic medical recommendations may be superior on the whole, their unique situation could best be appreciated by a human. In research directly comparing the relative amount of advice taking, Onkal and colleagues (2009) found that participants took more advice from a human expert as opposed to an algorithmic source when divergent pieces of advice were presented simultaneously. Overall, a robust literature has demonstrated many instances of algorithmic aversion, in sharp contrast to work documenting algorithmic appreciation, leading to questions of when one might confront algorithmic aversion or appreciation.

Who Prefers Algorithms and Why: Cognitive Style and the Mechanisms of Accuracy, Impartiality, and Objectivity

Whereas there is an abundance of research focusing on the features of the algorithm that cause algorithm aversion, there is very little research on characteristics of decision makers that predict algorithmic appreciation or aversion. Zhang and Dafoe (2019) demonstrated that greater understanding of algorithms and computer science education is predictive of support for algorithmic development. Logg and colleagues (2019) found that more numerate decision makers showed greater algorithmic appreciation in forecasting and weight estimation tasks. These results, in line with Yeomans and colleagues' (2019) finding that greater understanding (through increased transparency and information about the algorithm's process) leads to more algorithmic appreciation, offer some initial clues into traits that would lead some individuals to show a generalized tendency towards (versus against) algorithmic advice.

Despite these seemingly different factors, these papers suggest that individuals who are able to take the time to understand how the algorithms work are more convinced of their efficacy— whether it be through a stronger sense of numeracy, prior experience with computer science, or having explanations of their processes. If individuals who are more likely to spend time thinking about the process of algorithms are more likely to prefer them, then it stands to reason that those who generally tend to think more – the *cognitive style* of engaging in purposeful deliberation -- would be more likely to prefer algorithms also. Cognitive style is also highly correlated with education (Pennycook et al., 2012), with more deliberative individuals tending to advance further educationally. Finally, cognitive style has a mixed relationship with numeracy, with some pointing to the in-built nature of numeracy within measures of cognitive style (Sinayev & Peters, 2015), while others argue for its uniqueness (Sirota & Juanchich, 2011; Liberali et al., 2012). In the next section, we explore a range of theoretical connections between cognitive style and algorithmic aversion, leading to our studies that systematically test these relationships.

Cognitive Style

Reflective (versus intuitive) cognitive style is an individual's tendency to suppress intuitive thoughts and subsequently engage in deliberate, intentional thinking. Kahneman's (2011) dual systems theory contrasts the deliberate and "slow" System 2 thought processes to automatic, heuristic-based, "fast" System 1 cognitive processes. Multiple overlapping typologies and labels exist to separate out these two processes (for reviews, see Evans & Stanovich, 2013, or Hayes & Allinson, 1994). Although individuals use both System 1 and System 2 thinking at different times and with different tasks, each person has a cognitive style or a natural preference to rely on one versus the other. Cognitive style is also a better predictor of performance on a wide sample of tasks than measures of cognitive ability, thinking disposition and executive functioning (Toplak et al., 2011). Cognitive style was popularized by Shane Frederick (2005) with the creation of the three item Cognitive Reflection Test ("CRT"). The CRT is a parsimonious measure of which system is dominant that can be as predictive as multiple hours of combined cognitive tests, such as the Wonderlic Personnel Inventory, SAT scores, and the Need for Cognition scale.

Deliberative thinkers are more likely to prefer algorithms as compared to intuitive thinkers because they are likely to see algorithms as more credible. Scholars who study source credibility decompose credibility into three key features: accuracy, objectivity, and impartiality (Pornpitakpan, 2004). Perceived accuracy is defined as the subjective assessment of the frequency and severity of errors (Maier, 2005). Perceived objectivity is defined as how biased (versus unbiased) a source appears (Jacobson, 1969), whereas perceived impartiality is defined as the extent to which one perceives the information given serves the best interests of the advisee, even at the cost of the advisor's own needs (Neu et al., 2011).[3] We propose that these advisor features—accuracy, objectivity, and impartiality—serve as three mechanisms driving the difference in algorithmic aversion between individuals with varying cognitive styles, and we offer several theoretical reasons for why more deliberative individuals would be expected to believe that algorithmic advisors are more credible.

[3] In line with the literature, these three advisor features were also referenced by participants in our pilot study organically through open-ended text responses that justified their advisor preference. See pilot study.

Prior research gives some insight that suggests that individuals may believe that algorithmic advisors would be more accurate than human advisors, and these perceptions of accuracy may then produce algorithmic appreciation. Factually, algorithmic models have been found to be superior to human judgment across a wide range of settings due to their ability to process large data sets that would be otherwise incomprehensible to individuals (Metawa, 2017; Yu, 2017) and to avoid cognitive biases that can lead to inaccurate decisions (Blohm et al., 2020; Li et al., 2020). Deliberative (versus intuitive) thinkers believe in science more (Gervais, 2015) and use probabilities and numbers more (Liberali et al., 2014; Mastrogiorgio & Petracca, 2014), and therefore may better incorporate the aforementioned factual realities around algorithmic superiority and thus perceive algorithms to be more accurate, which then drives algorithmic appreciation.

Further, individuals who are more cognitively deliberate may believe that algorithmic advisors are more objective, i.e., less influenced by emotions and opinions than human advisors, which should translate into greater algorithmic appreciation. This is because, for cognitively deliberate individuals, the processes by which algorithms work, i.e., based on analytical reasoning that appears more objective, is likely of greatest salience. Indeed, the defining feature of those who are cognitively deliberate is the analytical process by which they think (Frederick, 2005). In contrast, cognitively intuitive individuals may attend more to the inputs associated with algorithm functioning because of their tendency to pay attention to inherent features as opposed to extrinsic factors. For example, theorists posit that most intuitive thinking focuses on inherent features that lead to purchase decisions (e.g., there is something inherently feminine about the color pink) instead of the extrinsic processes (e.g., marketing campaigns by clothing manufacturers influence the gendering of colors) that lead to preferences (Cimpian & Salomon, 2014). In the case of algorithmic advisors, intuitive individuals might focus more on the output from the advisor and miss the benefits to objectivity from the algorithm's process. As a result of deliberative thinkers' focus on the process that algorithms use, they may see algorithms as more objective than do their cognitively intuitive counterparts.

Lastly, individuals of varying cognitive styles may differentially demonstrate algorithmic aversion due to their differing perceptions of advisor impartiality. Cognitively intuitive individuals are more susceptible to disinformation, conspiracy theories, and 'bullshit' (Bago et al., 2020; Bronstein et al., 2019; Pennycook & Rand, 2019; Pennycook et al., 2015). Given the pervasive societal tropes of malevolent artificially intelligent technologies (Elsbach & Stigliani, 2019), intuitive individuals could suspect the algorithmic advisors' intentions and see them as less impartial. These concerns about ulterior motives could lead intuitive individuals to believe that the algorithmic advisors would place the interests of technology (or the technology's developer) above their own, leading to algorithmic aversion. Cognitively deliberative individuals, however, are less susceptible to these tropes (Toplak et al., 2011) and thus may recognize that a human advisor could easily prioritize the advisor's own gain over providing the best advice, and thus may prefer algorithmic advisors for their impartiality. Accordingly, we predict that deliberative individuals would still be more likely to recognize the benefits of impartiality that algorithmic advisors provide.

Overview of Predictions

Given the theory just reviewed, we predict that deliberative cognitive style will be related to reduced algorithmic aversion compared to individuals with a more intuitive cognitive style

(H1). We posit that this will be due to enhanced perceptions of advisor accuracy (H2a), objectivity (H2b), and impartiality (H2c), such that individuals with deliberative cognitive style (high CRT) will perceive greater accuracy, objectivity, and impartiality from algorithmic (versus human) advisors.

Overview of Studies

In order to examine the relationship between cognitive style and relative preference for input from an algorithmic advisor versus a human advisor in a realistic scenario, we asked participants to interact with either a human or algorithmic advisor through a chat textbox. In Study 1, we examined the correlations between cognitive style and preferences for an algorithmic versus a human advisor in a financial decision domain, controlling for key variables such as comfort with technology, social anxiety, and personality traits. Although the focus of this study was on cognitive style for the reasons stated above, justifications for the control variables included also precede each study. We replicated the relationship between intuitive cognitive style and algorithmic aversion even after controlling for self-perceived intelligence in Study 2. In Study 3, we extended the relationship beyond the financial decision domain by replicating the relationship between cognitive reflection and algorithmic aversion across multiple decision domains, including healthcare management and employee hiring decisions. Finally, in Study 4, we examined potential mechanisms that mediate the relationship between intuitive cognitive style and algorithmic aversion; we found that perceptions of advisor accuracy, and to a lesser extent advisor impartiality, mediate the relationship. To rule out alternative explanations, all studies in this paper controlled for demographic characteristics of age, gender, race, employment, political affiliation, annual household income, and educational attainment.

Although we retained other control variables as much as possible throughout the paper, we removed previously nonsignificant variables when later studies ran long, in order to prevent test fatigue. All studies were approved by the Committee for the Use of Humans as Experimental Subjects at the Massachusetts Institute of Technology. All participants provided informed consent.

Pilot Study

In order to measure the range of explicit justifications for a preference for an algorithmic (versus human) advisor, we asked a sample of 204 U.S. residents on Amazon Mechanical Turk to share their opinions on a hypothetical decision scenario. We told participants that we were interested in helping Americans develop their financial literacy and were seeking their opinion on tools that could help them navigate financial scenarios. Before they were introduced to the scenarios, we gave them the option to choose between an algorithmic or human advisor that were equally trained and knowledgeable and that would be able to answer any questions through a chat text box. On average, participants slightly preferred a human advisor over the algorithmic advisor (*M* = 2.65, *S.D.* = 1.4, where 1 = Prefer human advisor, 5 = Prefer algorithmic advisor). After choosing between the advisors, participants then gave an open-ended text response justifying their choice. The majority of responses stated ambivalence towards both advisors (35%). For those in favor of algorithmic technology (29%), respondents pointed to supposed AI performance advantages (16.5%) including accuracy, higher processing power, and recent developments in technological capabilities (7.5%), or a lack of emotional bias or ulterior motives (9%). In addition, respondents also expressed the desire to avoid human interaction (12.5%) due to social anxiety and/or embarrassment at their prior financial decisions. Those in favor of the human advisor (67.5%) emphasized humans' superior intellectual abilities (28.5%) such as the ability to customize advice (9.5%), draw from real-life experience (8.5%), or generalized trust in human thinking over machines (10.5%). Alternatively, some noted the value of a human's interpersonal abilities (39%), including the capacity to empathize and relate (10%), to communicate and answer questions organically (9%) and a vague sense of comfort in dealing or interacting with another human (20%). These open-ended text responses informed the selection of the control variables that we included in our studies.

Study 1

Study 1 established the main relationship between cognitive style (measured through the Cognitive Reflection Test, detailed below) and algorithmic appreciation. We tested this relationship in the domain area of financial advising, a domain area that has high external validity due to the prevalence of both human and algorithmic advisors. In order to test whether the effect was robust to relevant control variables, we also included participants' generalized comfort with technology, assuming that individuals who are more comfortable using technology in general would have less hesitation in their abilities to use an algorithmic advisor.

Method

Participants

508 U.S. residents were recruited from Amazon Mechanical Turk in return for market-rate compensation (M_{age} = 37.4, 53.6% women). We excluded 55 participants for not giving consent or for failing basic attention checks, such as "Please choose option 5 for this question", leaving 453 participants for analysis. This study was conducted in two separate waves investigating other hypotheses not reported in this paper. There were no significant differences

between the two waves, and as they followed the same basic procedure, we report the studies together. Data, materials, and analyses are available on OSF (https://osf.io/sz354/).

Procedure

We advised participants that the survey comprised a series of unrelated tasks.   We introduced the Cognitive Reflection Test as an "intelligence test" and we informed the participants that, even though the test was composed of only three questions, they should take their time answering it.

We then asked them to take part in a separate task involving a financial decision-making scenario. We asked participants to imagine themselves looking for advice on managing an investment portfolio of financial assets and that we would give them the opportunity to interact with a financial advisor over a text chat box.

Participants had two options for their advisor – an algorithmic advisor and a human advisor stated to be of equivalent ability, speed, and cost – to answer questions that they might have. The advisor, they were told, would ask them questions about their lifestyle, goals, and background and then tailor their advice on the investment portfolio to their answers. They were then provided a continuous sliding scale whereby they could choose the relative amount of advice from both the algorithmic advisor and the human advisor (more detail provided below).

After choosing the advisor, participants saw an error message that stated that the rest of the financial advice scenario could not be loaded, but that their work would still be compensated. As we were interested in advice seeking behavior measured by their choice on the sliding scale, we did not actually let the participants interact with either of the (hypothetical) advisors, nor did they receive any advice. The participants went through the remaining scales and answered demographic questions before being thanked and paid for their time.

Measures

Advisor preference. The focal outcome of advisor choice was self-reported on a sliding scale ranging from 0-100 (increasing in increments of 10; in later studies, this restriction was relaxed), with all advice coming from the algorithmic advisor being 0, all advice coming from the human advisor being 100, and an even split of advice from both being 50. The choice between human and AI sources of information is not binary; people can and do refer to multiple sources for advice. Therefore, we operationalized algorithmic aversion and algorithmic appreciation as the stated preference for the relative amount of input from a human and an algorithmic advisor. This provides a continuous variable that represents the real choices that people make as they navigate technology in their choice tasks.

Cognitive style. We measured cognitive style through the original Cognitive Reflection Test (CRT, Frederick, 2005). This three-item measure comprises questions with an intuitive (but incorrect) answer that must be overcome with deliberation in order to reach to the correct result. An example question is:

"A ball and a bat cost $1.10.

If the bat costs $1.00 more than the ball,

How much does the ball cost?"

In this question, $0.10 is the intuitive, but incorrect, answer, and $0.05 is the correct answer. The scale questions and answers are available in the online supplement. Participants were able to give any numerical response; we coded them as correct or incorrect and created a score of percentage of items answered correctly (0-100%) (Cronbach's α = 0.74). The percentage score allows comparison across all studies, some of which used a seven-item CRT version to be described later. Unless explicitly stated in the methods sections, CRT scores are based on the original three-item version.

Comfort with technology. We adapted a comfort-with-technology scale used in online-education settings (Rodriguez, Ooms, and Montañez, 2008) for a broader general audience. Respondents used the revised 8-item scale (Cronbach's α = 0.88) to rate their comfort (1 = *Very comfortable*, 5 = *Very uncomfortable*) accomplishing a variety of tasks involving technology (e.g., "Downloading and reading e-books," "Use social media to connect with a stranger," "Save and retrieve files in the cloud"). We reverse coded the summed score, so that higher numbers signify greater comfort with technology. The mean in our sample was 24.6 (*S.D.* = 5.8, *Min.* = 0, *Max.* = 30), suggesting a range of proficiency levels with technology.

Social anxiety. Since the pilot study revealed that social anxiety was a commonly cited reason for avoiding the human advisor, we measured social anxiety as a potential factor driving preference for an algorithmic (over a human) advisor. Socially anxious individuals may be more likely to avoid interacting with another human due to a concern about being judged in a social interaction.

We gave participants the short form of the Social Interaction Anxiety Scale (SIAS) that is widely used by clinicians (Fergus et al., 2014). This six-item scale (Cronbach's α = 0.86) has been validated in both clinical and non-clinical samples to be predictive of interpersonal functioning, and has shown convergence with other measures of social anxiety (Fergus et al., 2012). Overall, participants scored a mean of 17.6 (*S.D.* = 5.8, *Min.* = 6, *Max.* = 30) out of a potential 30, with higher scores denoting greater social anxiety.

BFI-10 Personality index. To rule out the possibility that individual differences in personality drive desires to engage in a new technology, we also included the ten-item version of the Big Five Inventory Personality Index. As the pre-eminent personality research inventory, the Big Five traits have decades of research demonstrating their external validity and factorial uniqueness (John & Srivastava, 1999; McCrae & Costa, 1987). We had expected extroverts to prefer interaction with another human instead of a task-focused algorithmic advisor, and that individuals high on openness to experience would be more interested in engaging with a novel algorithmic advisor. We had no expectations for conscientiousness, neuroticism, or agreeableness.

To control for differences in openness to experience and extroversion as predictors of preference for an algorithmic advisor, we gave participants the short-form (10 items) of the Big Five Inventory (Rammstedt & John, 2007), which has similar reliability and predictive power as the longer version. Reliability in our sample was adequate for all five traits: openness to experience (Cronbach's $\alpha = 0.61$), neuroticism ($\alpha = 0.78$), agreeableness ($\alpha = 0.41$), conscientiousness ($\alpha = 0.62$), and extroversion ($\alpha = 0.67$).

Confidence in financial literacy. Lastly, given the financial advice seeking scenario, we included a measure of confidence in financial literacy in case individuals who were highly confident in their financial ability might be more likely to dismiss the potential benefits of talking to another person and thus be more receptive to a novel advice source from an algorithmic advisor. Alternatively, individuals confident in their financial literacy may feel less shame discussing their situation with another person and therefore show greater preference for a human advisor.

Participants self-reported their confidence in their financial literacy on a one-item question using a 5-point Likert scale (1 = *Very confident*, 5 = *Not at all confident*). For ease of comprehension, we reverse scored the scale so that higher scores denoted greater confidence in financial literacy (*M* = 3.4, *S.D.* = 1.0, *Min.* = 1, *Max.* = 5).

Results

Our analyses examined the hypothesized relationship between cognitive reflection and algorithmic aversion (measured by preference for human versus algorithmic advisor). We also examined whether the relationship was robust to the inclusion of control variables, including comfort with technology, social anxiety, confidence in financial literacy, personality traits, and demographic characteristics.

Algorithmic Aversion

Participants slightly preferred more advice coming from a human advisor over an algorithmic advisor (*M* = 55.7%, *S.D.* = 25.6%, *Min.* = 0%, *Max.* = 100%; where 0% = all advice from algorithmic advisor, and 100% = all advice from human advisor). The complete distribution in Figure 1 shows that participants most frequently answered 50, suggesting equal proportions of advice from the human and algorithmic advisors.

Cognitive Style and Algorithmic Aversion

The average score on the CRT for the participants in this sample was 48.8% (*S.D.* = 40%), indicating that participants correctly answered a little less than one-half of the questions. For comparison, the average CRT score collected across 3,000+ individuals in a diverse range of locations was 1.24 or 41% (Frederick, 2005), slightly lower than our online sample. The distribution of standardized scores, as shown in Figure 2 below, is bimodal, with many participants getting all or none of the questions right.

When we subjected the data to ordinary least squares regression (Model 1a in Table 2) without controls, we see, consistent with our hypothesis, that preference for advice from a human advisor decreased as cognitive reflection increased ($B_{cognitive\ reflection} = -13.7$, *S.E.* = 2.9, $p < 0.001$, 95% CI: [-19.49, -7.91]).

Control Variables

Many of the control variables were significantly related with advisor preference, although not all were in the direction we expected (see Table 2). As expected, more extraverted individuals preferred greater advice from the human (versus algorithmic) advisor. Openness was not related to advisor preference. The only other BIG-5 Personality trait that was significantly related to algorithmic preference was conscientiousness, such that more conscientious individuals preferred a greater amount of advice from the human advisor.

Although we had expected that individuals would feel embarrassed by their low financial literacy when judged by a human advisor and thus prefer more advice from an algorithmic advisor, we found that individuals who were more confident in their financial literacy actually preferred more advice from an algorithmic advisor. Also contrary to our expectations, socially anxious individuals did not shy away from human advisors, but rather showed significantly more algorithmic aversion than those lower on social anxiety, albeit only slightly, with an increase in one point on the social anxiety scale resulting in an increase of less than 1% in the additional amount of human advice sought.

Of the demographic variables, only being unemployed (and searching for employment) or retired were related to algorithmic aversion and being politically Independent was related to algorithmic appreciation. However, even when all these control variables were included in the models, the effect of CRT was consistent and unreduced in magnitude. All variables across both models had Variance Inflation Factors smaller than 5, mitigating any concerns of multicollinearity between cognitive style and the other controls.

Discussion

As hypothesized, we found a robust relationship between cognitive style and algorithmic aversion, such that cognitively intuitive individuals showed greater algorithmic aversion than their deliberative counterparts. Even after controlling for participants' generalized comfort with technology, personality traits, confidence in financial literacy, social anxiety, and demographic characteristics, the relationship between cognitive style and algorithmic advisor aversion was robust.

The unexpected relationship between algorithmic appreciation and confidence in financial literacy caused us to reflect. It may be that confidence in financial literacy reflects one aspect about confidence in general. More confident or self-efficacious individuals may rely less on the help of another person, but feel able to independently interpret the advice given from an algorithmic source. In order to investigate further, we tested two alternate measures of self-competence in Study 2: growth mindset (Dweck, 1986), or beliefs about one's ability to learn new skills, and self-perceived intelligence (De Keersmaecker et al., 2017).

Study 2

We sought to replicate the relationship between cognitive style and algorithmic aversion to see whether it was robust to beliefs about individual competence, as measured in two ways: self-rated intelligence and growth mindset. Individuals who have a generalized belief in their intellectual superiority, not limited to financial literacy, may feel that another human advisor would be less intelligent than themselves and thus would derive fewer benefits from the advice from the human advisor. As elaborated in the discussion for Study 1, individuals who are confident in their self-efficacy may also feel more comfortable interpreting advice from an algorithmic source, and thus would show algorithmic appreciation. To ensure that our paradigm was robust against concerns that individuals who felt smarter than other people would rely less on advice from the human advisor, we asked participants to rate their own intelligence in comparison to other individuals and controlled for this factor when investigating the focal relationship of cognitive style and algorithmic aversion.

In addition, we chose growth mindset as a reliable and well-established measure of an individual's beliefs about their intellectual abilities (Dweck, 1986). Individuals with a fixed mindset, compared to those with a growth mindset, believe that cognitive capacities are innate, and may therefore believe that they are less able to learn how to use a novel technology (as presented by the algorithmic advisor), and thus show algorithmic aversion. Because cognitive intuition captures the tendency to rely on automatic and heuristic-based processing, individuals who are low on cognitive reflection may tend toward a more fixed mindset, and be less likely to interact with a novel technology that would involve developing additional skills. Alternatively, a fixed mindset may reflect a belief in inherent limitations on human reasoning, and therefore greater motivation to rely on the algorithmic advisor. Given the lack of research connecting cognitive style with mindset (see review by Rattan and Georgeac, 2017), we took the opportunity to examine this relationship.

We also sought to explore the robustness of the relationship between cognitive style and algorithmic aversion by using alternate measures of cognitive reflection and social anxiety. Given the surprising relationship between social anxiety and algorithmic aversion that seemingly contradicted results from the pilot study, we included an alternative measure of social anxiety commonly used by clinicians (Heimberg et al., 1999) to verify the directionality of the effect. And, although the three-item CRT is a widely used measure of cognitive style, we sought to replicate the focal effect with an alternative version that relies less on numeracy, detailed below.

Methods

Participants

242 U.S. residents were recruited from Amazon Mechanical Turk in return for market-rate compensation (M_{age} = 35.9, 52.7% women). Thirty-five participants were dropped from the analyses for failing basic attention checks such as "Please choose Five for this option", leaving 207 complete responses.

Procedure

We used the same paradigm as in Study 1, but with revised scales as described below.

Measures

Cognitive Style (Seven-Item CRT). In order to ensure that our results in Study 1 were not limited to one measure of cognitive style, we used a modified seven-item version of the original Cognitive Reflection Test (CRT) that is less focused on numerical calculations yet is reliable and valid (Pennycook & Rand, 2019). This combined measure consisted of three reworded items from the original CRT that were computationally equivalent but semantically different, e.g., estimating the size of a patch of mold on a loaf of bread (modified) instead of estimating the lily-pad coverage of a pond (original) (Shenhav et al., 2012) and four non-numeric items that had a false lure answer (e.g., "Emily's father has three daughters. The first two are named April and May. What is the name of the third daughter?") that needed to be suppressed in order to arrive at the correct answer (e. g., Emily, not June) from Thomson and Oppenheimer (2016). The seven-item measure had acceptable reliability, Cronbach's α = 0.78. Scores could range from zero (indicating all questions were answered incorrectly and thus the least amount of cognitive reflection) to seven (indicating the highest amount of cognitive reflection), which were then standardized to the 0-100% scale used in Study 1 for ease of comparison across studies. Overall, participants answered 49.1% of the questions correctly (*S.D.* = 30%, *Min.* = 0%, *Max.* = 100%), very similar to the scores on the 3-item version in Study 1.

Alternate measure of Social Anxiety. We included the Liebowitz Social Anxiety Scale (Liebowitz, 1987), commonly used in clinical settings (Heimberg et al., 1999). This scale consists of 24 situations that participants rate on how much they avoid them. We rephrased the question to read "How much do you try to avoid the following situations?", and participants replied for each situation on a 0-4 Likert scale ranging from "Never" to "Usually (67-100% of the time)" in the past week. Scores above 50 indicate some form of social anxiety, with scores in the 50-65 range indicating moderate social anxiety, and scores of 95-100 indicating very severe anxiety. In this sample, the mean score was 31 (*S.D.* = 15, *Min.* = 0, *Max.* = 72), suggesting that, on average, participants did not demonstrate clinically severe social anxiety, but a broad range of anxiety levels were present in the sample, with 12.6% of the sample demonstrating moderate to severe social anxiety.

Self-Perceived Intelligence. We asked one question to measure self-perceived intelligence: “Compared to the rest of the population, how intelligent do you consider yourself?” Participants could answer on a sliding scale that ranged from “Less intelligent than everyone else” (0) to “More intelligent than everyone else” (100), with the midpoint labelled as “More intelligent than half the population” (50). This question was asked immediately before taking the Cognitive Reflection Test, with no forewarning of the test in advance. On average, individuals rated themselves to be smarter than 66.6% (*S.D.* = 14.7, *Min.* = 23, *Max.* = 100) of the population.

Mindset. To measure participants’ lay theory of cognitive malleability, we gave them a three-item measure of Intelligence Mindset (Dweck et al., 1995), for example, “You have a certain amount of intelligence and you can’t really do much to change it”. Responses were on a 6-point scale from 1 = “Strongly Agree” to 6 = “Strongly Disagree”, which were reverse scored for ease of comprehension. The mean score on the measure of Mindset was 6.6 out of a potential 18 (*S.D.* = 4.2, *Min.* = 0, *Max.* = 15), suggesting on average that participants “mostly agreed” with statements implying a malleable mindset.

Results

The relationship between cognitive style and algorithmic aversion replicated with the more comprehensive measure of cognitive style, with more cognitively intuitive individuals demonstrating greater algorithmic aversion than deliberative individuals ($B_{cognitive\ reflection}$ = -19.1, *S.E.* = 5.8, *p* = 0.001, 95% CI: [-30.5, -7.76]; see Model 2a in Table 4). This effect was robust to the inclusion of the control variables of social anxiety (using the modified Liebowitz Social Anxiety Scale), self-perceived intelligence, and growth mindset beliefs (see Model 2b), as well as with and without demographic controls (see Model 2c). All variables across all models had Variance Inflation Factors smaller than 5, addressing concerns of multicollinearity.

Control variables

We found no relationship between self-perceived intelligence and CRT score, possibly due to so many people rating their intelligence as above average, a version of the 'Lake Wobegon'[4] effect. There was a modest relationship between self-perceived intelligence and advisor preference ($B_{Perceived\ Intelligence}$ = 0.24, *S.E.* = 0.12, *p* = 0.05; See Model 2b, Table 4), such that individuals who perceived themselves as more intelligent very slightly preferred more advice from the human advisor, but this effect disappeared with the inclusion of demographic controls (Model 2c).

In contrast to our expectations that individuals with a growth mindset would be more open to a novel technology or that socially anxious individuals would prefer more advice from an algorithmic (versus human) advisor, we found no significant relationships between mindset or social anxiety and advisor preference (see Models 2b and 2c in Table 4).

Discussion

The results from Study 2 confirmed that the relationship between cognitive style and algorithmic aversion replicated with a new sample and was robust to perceptions of self-competence, as measured in two ways: growth mindset beliefs and self-perceived relative intelligence. Those who generally believe they can learn and master new concepts could just as easily prefer learning from a human advisor as an algorithmic advisor. It could also be that

[4] For readers unfamiliar with the reference, 'Lake Wobegon' is a fictional town, envisioned by popular radio personality Garrison Keillor, where "all the women are strong, all the men are good-looking, and all the children are above average".

algorithmic advisors are so commonplace that the barrier to learning how to use its advice was too low to be affected by growth mindset beliefs. Similarly, those who generally felt smarter than most other individuals did not exhibit a different preference for advice from a human or algorithmic advisor.

In addition, we were able to rule out social anxiety as an explanation for algorithmic appreciation, since neither clinically-validated measure used in the studies was related to greater preference for algorithmic advice. In fact, there was a modest preference for socially-anxious respondents to want more human advice in Study 1, but not in Study 2. Perhaps the format of a text box chat does not engage social anxiety, and a more intimate medium, such as phone call, video conference, or in-person interaction, would show different results. However, given that the pilot testing revealed social anxiety as an explanation for avoiding human interaction in the same chat text box format, we conclude that individuals may be largely unaware of the sources of their advisor preferences and offer social anxiety as a kind of hypothesis.

Lastly, we were able to confirm that the relationship between cognitive style and algorithmic aversion was not limited to scores from the original Cognitive Reflection Test and replicated with a less numeracy-based measure. Although not the main focus of their paper, Logg et al. (2019) identified numeracy as a secondary finding to their main discovery of generalized algorithmic appreciation. Although cognitive style has some relationship to numeracy, our finding that intuitive cognitive style is associated with algorithmic aversion replicates across both versions of the CRT and even with a new score calculated from only the four non-numeric items in the CRT ($r(242) = -0.28$, $p < 0.001$, 95% CI [-0.39, -0.16]).   This suggests that it is cognitive style, apart from numeracy, that predicts algorithmic aversion.

However, both Studies 1 and 2 are limited to the single decision domain of financial decision-making. Study 3 sought to extend the research to a variety of decision domains.

Study 3

Studies 1 and 2 established and replicated a robust relationship between cognitive style and algorithmic aversion, where more cognitively intuitive individuals prefer a greater amount of advice from human compared to algorithmic advisors, but solely within the financial decision domain. Previous research has established that algorithmic aversion varies by decision domain. When making medical decisions, individuals tend to exhibit algorithmic aversion (e.g., Promberger & Baron, 2006 and Longoni, Bonezzi, & Morewedge, 2018), as compared to making mathematical forecasts and quantitative estimates, where algorithms are preferred (Dietvorst & Bharti, 2020). Given that our paradigm focused on financial advice, individuals may have shown more openness to the algorithmic advisor than in a less quantitative domain.

Study 3 therefore sought to establish whether the relationship between cognitive style and algorithmic aversion depends upon decision domain. In addition to financial decisions, we chose three other decision-making domains where algorithmic advisors are already in use: healthcare management, employee hiring, and college admissions (see introduction for examples). Theoretically, although the research literature would lead us to expect differences in algorithmic aversion by domain, for example, more algorithmic aversion in healthcare settings (Longoni et al., 2019; Promberger & Baron, 2006), there was no reason to expect that the relationship between cognitive style and algorithmic aversion would differ by decision-making domain.

Methods

Participants

633 U.S. residents were recruited from Amazon Mechanical Turk in return for market-rate compensation (M_{age} = 38.1, 61.9% women). We eliminated 120 individuals, leaving 513 respondents behind, for the following reasons: 2 individuals did not consent, 15 discontinued the survey before completion, 15 failed a basic attention check surrounding the content of the decision-scenario, 27 failed a basic attention check about the advisors in the scenario, and 61 failed a basic attention check where they were asked to select a specific option within a Likert scale.

Measures

Decision-Making Domain. We added three new decision domains to the original financial investment decision task: College Admissions, Employee Hiring, and Healthcare Management (full survey materials are available to download at https://osf.io/sz354/). We randomly assigned participants to one of the four decision domains, for which they were asked to imagine themselves as making a decision and needing additional advice (College admissions: *You are helping make college admissions decisions and are looking for advice on how to evaluate applicants*; Employee hiring: *You are in the HR department and are looking for advice on how to evaluate applicants for hiring*; Healthcare Management: *You have been diagnosed with a health condition and are wanting advice on how to manage it*). We intentionally left the scenarios vague (with no mention of the inputs that would be used by the advisors) so that each individual could interpret the decision as being subjective or objective as they wished. We did this intentionally so that the participants were able to project onto the decision scenario whatever information they wished about the domain area. The survey varied only in this task introduction, and otherwise provided the same information about the algorithmic and human advisors.

Results

As shown in Table 6, although decision domains differed significantly on algorithmic aversion, cognitive style maintained the same relationship with advisor preference. Across the four decision-making domains, healthcare management was significantly different from the reference category domain of financial decision-making in predicting preference for algorithmic versus human advice ($B_{Healthcare\ Domain}$ = 12.108, *S.E.* = 3.0, $p < 0.001$, 95% CI: [5.93, 17.98]), with individuals preferring on average 12% more advice coming from a human (versus algorithmic) advisor than individuals making decisions in the other domain areas. Despite this domain effect, individuals seeking advice for healthcare management decisions still showed the relationship of lower CRT scores with less advice sought from the algorithmic advisor (*r*(128) = -0.17, *p* = 0.048). There were no interaction effects of domain area with cognitive style in predicting algorithmic aversion, as shown in Models 3d and 3e (with demographic variables) in Table 6, suggesting that the relationship between cognitive style and algorithmic aversion was robust across decision domain. Across all models, regardless of controls, the relationship between cognitive analytical style and advisor preference remained substantively unchanged (in the full model: *B* = -15.406, *S.E.* = 5.4, $p < 0.01$, 95% CI: [-26.3, -4.52]; see Table 6 for all coefficients).

Discussion

The results from Study 3 confirmed that the relationship between cognitive style and algorithmic aversion held across a variety of decision-making domains. This finding is one of the first generalizable individual-level differences that predicts algorithmic aversion across multiple decision areas.

Our findings reinforce prior research showing domain differences in algorithmic aversion. It is noteworthy that individuals in the healthcare management decision domain preferred significantly more human advice than those in any of the other decision domains. Although it was not the goal of this research to explain domain difference, it is still unclear why healthcare management would be different. Previous research has pointed to uniqueness neglect, where individuals believe that while algorithmic models are generally good at predicting estimates on average, they expect that human advisors can better tailor and customize advice for their own unique situation (overestimating their uniqueness) (Longoni et al., 2019). Other work showing algorithmic aversion in a healthcare setting suggests that people are more able to pass responsibility for a bad decision to a doctor rather than an equally capable algorithmic recommender tool (Promberger & Baron, 2006). More recently, Dietvorst and Bharti (2020) posed that individuals prefer algorithmic recommendation systems when the decision domain is more epistemically uncertain.

However, these reasons do not seem restricted to the medical decision domain. An individual could also have concerns that something unique about their financial situation may be overlooked by an algorithmic system, or an individual may take comfort after a college rejection by blaming a college admissions counsellor (as opposed to a blameless algorithm). Similarly, a healthcare setting is not inherently less uncertain than financial investing, or the eligibility of a high school senior for college admissions. Thus, explaining heightened algorithmic aversion in the healthcare decision domain will require additional future research.

Overall, we have found evidence for a robust and replicable relationship between cognitive style and algorithmic aversion in the studies so far. The remaining question lies in the mechanism by which individuals of varying cognitive styles choose between algorithmic and human advisors. In order to test the directionality of the effect, and understand the reasons for exactly why individuals of varying cognitive styles show differing amounts of algorithmic aversion, we tested the perceptions that deliberative (versus intuitive) individuals held of both kinds of advisors and compared them to each other.

We also investigated the inclusion of the relevant control variable of prior experience with AI in Study 4. Individuals with many bad experiences or very little experience at all with AI may be less likely to opt for advice from an algorithmic advisor, whereas individuals familiar with AI may have more favorable pre-existing outlook towards algorithmic tools. The mere exposure effect, or familiarity principle (Bornstein & D'Agostino, 1997), suggests that familiarity with a technological product of system may increase preference. Thus, individuals with more prior experience with AI may be more likely to prefer an algorithmic advisor, and this may overshadow the effect of cognitive reflection. In order to more convincingly show the relationship between cognitive style and algorithmic aversion, we included prior experience with AI as a control variable to show that it was robust to this relevant variable.

Study 4

Study 4 sought to understand the mechanism by which an individual's cognitive style might influence algorithmic aversion. Our approach was to consider an advisor, whether human or algorithmic, as a source of information, and to therefore draw on the source credibility research literature to examine three key features of information sources: accuracy, objectivity, and impartiality (Pornpitakpan, 2004). We hypothesized that individuals with a more analytical cognitive style may evaluate algorithmic advisors as more credible sources of information because of one or more of these features: (1) perceptions of advisor *accuracy* (a focus on performance outcomes), (2) *objectivity* (not being influenced by personal feelings or opinions in representing facts), and (3) *impartiality* (putting the advisee's needs and interests ahead of the advisor's needs and interests).

In Study 4, we used the same paradigm as in the previous studies but collected additional perceptions of the advisors as ratings of accuracy, objectivity, and impartiality. We also sought to improve the measure of prior experience with AI to ensure that we are controlling the nature of prior experience with AI when assessing the relationship between cognitive style and algorithmic aversion.

Methods

Participants

1,198 U.S. residents were recruited from Amazon Mechanical Turk in return for market-rate compensation (M_{age} = 38.6, 57.3% women). Four participants attempted to take the survey twice, 76 failed a basic attention check asking about the decision-making area in the scenario (out of a four multiple-choice question), 18 failed an attention check about the kinds of advisors in the scenario (also out of a four multiple-choice question), and 48 did not complete the survey, leaving 1052 responses. This survey was conducted in four waves investigating separate hypotheses not covered in this paper, and since the recruitment, participant population, and procedures used were the same, we condensed them into one study with increased power.

Procedure

We used the same paradigm as the studies above, but one-half of the participants (632, 60%) took the CRT immediately after choosing between the two advisors and giving a written rationale for their decision, as in Study 3, whereas the others took the CRT first. Order effects are discussed below. After choosing between the advisors, all participants answered six questions about their perceptions of the advisors, as detailed below.

Measures

We assessed cognitive style using the 3-item CRT as in prior studies, and created new measures of advisor perceptions.

Advisor Perceptions. We asked participants to give ratings of both advisors (human and algorithmic) on 5-point Likert scales for accuracy, objectivity, and impartiality. For accuracy, the Likert scale ranged from accurate (1) to inaccurate (5). Objectivity was defined for participants as "not influenced by personal feelings or opinions in considering and representing facts" and the scale ranged from objective (1) to subjective (5). For impartiality, we asked “How much do you think these advisors put your needs and best interests first?”, where 1= “Put **my** needs ahead of their [the advisor’s] needs” and 5 = “Puts **their** needs ahead of my needs”. For ease of interpretation, the responses for each feature were reverse scored so that larger numbers represent greater accuracy, objectivity, or impartiality. The majority of the six perceptions were significantly, positively correlated, with a few exceptions, as seen in Table 7.

Prior experience with AI. We asked participants to recall their previous experiences with AI across three task areas: accomplishing a task, getting information, and getting tailored advice using artificial intelligence. Participants answered each of these items on a 5-point Likert scale (1 = *Not at all effective,* 5 = *Very effective*, with an option of “Not Applicable”). We summed together their answers to give a score measuring the quality of their prior experiences using AI (Cronbach's $\alpha = 0.8$). Participants in this sample had generally positive experiences, with their average rating across the three categories suggesting somewhat to very effective prior experiences with AI agents ($M = 4.0$, $SD = 0.99$). Prior experiences with AI and cognitive style are not significantly correlated ($r = 0.03$, $p = 0.3$). No respondents chose the option of "Not Applicable".

Results

The relationship between cognitive analytical style and advisor preference was replicated, with higher CRT scores associated with preference for advice from the algorithmic advisor ($B_{cognitive\ reflection} = -11.7$, *S.E.* = 1.9, $p < 0.001$, 95% CI: [-15.5, -8.0]; See Model 4a in Table 8). Quality of prior experience with AI was not significantly related to advisor preference (*p*s = 0.8, see Models 4d and e in Table 8). Similarly, there was no difference in taking the CRT before or after deciding between advisors (*p*s > 0.1, see Models 4b and c in Table 8).

Advisor Perceptions

In order to assess whether advisor perceptions of accuracy, objectivity, and impartiality varied with cognitive style, we ran a multivariate analysis of variance of the six advisor features along with cognitive style. As expected, the perceptions of the algorithmic advisors significantly varied by cognitive style, Pillai's Trace = 0.01, $F(3, 1048) = 4.94$, $p = 0.002$. Post-hoc analyses revealed that while cognitively intuitive and deliberative individuals rated algorithmic advisors significantly differently on objectivity ($F(1, 1050) = 10.9$, $p < 0.001$) and impartiality ($F(1, 1050) = 6.5$, $p = 0.01$), the difference between perceptions of algorithmic advisor accuracy was only marginally significant, ($F(1, 1050) = 3.12$, $p = 0.078$). Similarly, the perceptions of the human advisors significantly varied by cognitive style, Pillai's Trace = 0.02, $F(3, 1048) = 6.4$, $p < 0.001$. Post-hoc analyses revealed that there was no significant difference due to cognitive style in ratings of human advisor objectivity ($F(1, 1050) = 0.1$, $p = 0.83$). However, cognitively intuitive and deliberative individuals rated human advisors significantly differently on accuracy ($F(1, 1050) = 11.1$, $p < 0.001$) and impartiality ($F(1, 1050) = 6.5$, $p = 0.01$).

Mediation

Using Baron and Kenny's (1986) steps for multiple mediation, we examined perceptions of advisor accuracy, objectivity, and impartiality for each advisor as mediators for the relationship between cognitive style and algorithmic aversion. As shown in Figure 2 below, both perceptions of human advisor accuracy and algorithmic advisor impartiality mediated the relationship between cognitive style and algorithmic advisor preference.

Following the steps for multiple mediation analysis from Hayes (2012) in using the "lavaan" package in R, we calculated that perceptions of accuracy of the human advisor mediated 20.5% of the variance ($B = -0.10$, *S.E.* = 0.03, 95% CI [-0.16, -0.04], $p = 0.001$), while accuracy of the algorithmic advisor was only marginally significant ($B = 0.06$, *S.E.* = 0.04, *95% CI* [-0.13, 0.01], $p = 0.08$), based on 5,000 bootstrapped samples. Perceptions of objectivity of the algorithmic and human advisors did not mediate the relationship ($p = 0.8$ and 0.8, respectively). For perceptions of impartiality, ratings of the algorithmic advisor mediated 4.0% ($B = -0.02$, *S.E.* = 0.01, 95% CI [-0.04, -0.00], $p = 0.04$), whereas perceptions of the human advisor did not mediate the relationship ($p = 0.6$). The inclusion of both significant advisor perceptions (human advisor accuracy and algorithmic advisor impartiality) resulted in mediation[5], with cognitive style providing a weaker predictor of advisor choice when the two perceptions were taken into account. This means that perceptions of advisors systematically vary for individuals of differing cognitive styles, such that deliberative individuals believe that human advisors are less accurate than do their intuitive counterparts, and deliberative individuals believe that algorithmic advisors are more impartial than intuitive thinkers perceive.

Discussion

The results from Study 4 shed light on why individuals of varying cognitive styles may show algorithmic aversion. Cognitively intuitive individuals believe human advisors to be superior in accuracy and algorithmic advisors to be less impartial, as compared to deliberative individuals who believe human advisors to be less accurate and algorithmic advisors to be more impartial. These beliefs mediated the relationship of cognitive style with algorithmic aversion, with differences in the perceptions of human advisor accuracy accounting for the majority of the mediation effect. Unexpectedly, perceptions of objectivity for either the algorithmic or human advisor did not mediate the effect between cognitive style and algorithmic aversion. One possible reason for this could be due to the unexpected consensus between individuals of varying cognitive styles in their assessment of human advisor objectivity, where the majority of respondents rated the human as less objective than the algorithmic advisor. The absence of a difference in objectivity is not as surprising as the significant divergence in perception of algorithmic advisor impartiality; Since algorithms are immune to bias from extraneous

[5] The term ‘full’ and ‘partial’ mediation have been subject to controversy and thus we use these terms cautiously in this paper. Rucker et al. (2011) give a cogent summary of the reasons why full mediation can impede theory development.

situational factors like human emotion, it gives more credence to our theory that intuitive individuals may be focused more on the output than the process used by algorithmic advisors. Overall, these results show promise for many future studies exploring this individual difference of cognitive style in greater detail.

General Discussion

In this paper, we find a robust relationship between cognitive style and preference for advice from an algorithmic versus human advisor, with more intuitive individuals preferring relatively more advice from a human (versus algorithmic) source. This relationship was robust to using the original 3-item CRT or a revised 7-item CRT that was less dependent on numerical questions, controls for the BIG-5 personality traits, social anxiety, comfort with technology, prior amount and quality of experiences with AI, cognitive mindset, self-perceived intelligence, self-reported financial literacy (for the financial decision-making scenarios), decision domains, and demographic variables. This work is one of the first papers to our knowledge to specifically focus on individual level differences as drivers for algorithmic aversion.

We also add to the sparse literature on the relationships between basic demographic characteristics and algorithmic aversion. In addition to the focal relationship between cognitive style and algorithmic aversion, we also identified some intriguing individual level differences that were related, albeit less strongly, to algorithmic preference, such as political affiliation, employment status, financial literacy, and social anxiety, as well as the personality traits of extraversion and conscientiousness, that could inspire further investigation in future work.

This work also contributes to the technology acceptance literature in three main ways: turning the focus on the user, underscoring the complexity between anthropomorphism and algorithmic aversion, and modernizing the kinds of technology studied. First, the existing literature has focused mostly on features of technical systems, rather than aspects of decision makers that affect technology acceptance (Lee et al., 2003). As well as identifying the differential way that individuals of varying cognitive styles perceive algorithmic advisors and their advice accuracy, this work has also shed light on several other individual level differences that may have surprising effects on algorithmic aversion: personality traits, as well as demographic variables of political affiliation. Second, whereas the technology acceptance literature has generally touted the benefits of humanizing non-human machines (Waytz, Heafner, & Epley, 2014), our work suggests a more complex relationship with anthropomorphism. Based on our results, we would expect that humanization of AI advisors would increase advice adoption for cognitively intuitive individuals, but would actually diminish the perceived accuracy of the AI advisor's recommendations for more deliberative individuals and thus decrease technology adoption. Lastly, this work modernizes the technology acceptance literature through the examination of artificially intelligent technologies, such as algorithmic advisors. The majority of the foundational work on technology acceptance models dates from the 1990s and focused on acceptance of printers and computers (Davis, 1989). As modern algorithmic agents can make subjective judgments more accurately than close friends and family members (Yeomans et al., 2019) and even provide emotional support (Lucas et al., 2017), our understanding of human-AI interaction must be correspondingly updated and enriched.

Our work also builds on the large body of research linking cognitive style and diverse behavioral outcomes, from voting behavior (Pennycook & Rand, 2019), and religious prejudice (Franks & Scherr, 2018) to overbidding at auctions (Sheremeta, 2018). Despite the importance of technology use in modern day life, hardly any work has examined the relationship between CRT and technology use (Barr et al., 2015; Vujic, 2017). Our work establishes an important link between cognitive style and preferences around artificially intelligent technologies, opening the door for work that experimentally manipulates cognitive style to affect technology acceptance[6].

Managerial Implications

For practitioners implementing novel AI agent-based features, these results suggest an easier adoption trajectory for target audiences naturally higher in cognitive reflection (e.g., non-religious, socially liberal, low testosterone individuals, Shenhav et al., 2012; Pennycook et al., 2016; Bahçekapili & Yilmaz, 2017; Deppe et al., 2015; Iyer et al., 2012; Nadler et al., 2017). For employers, this quick three-item measure of cognitive style can help assess a future hire's proclivity towards algorithmic or human advice sources, either in the form of actual advisors or in assistive programs. It also identifies a pre-emptive approach toward digital transformation, whereby organizations and managers can strategize on the best framing of a new digital product depending on the cognitive style of their audiences and tailor their approach accordingly.

Another practical discovery from our current work is the effect that the decision domain holds on algorithmic aversion. In line with other work that has documented algorithmic aversion in medical decision-making domains, we found that individuals prefer human advisors more than algorithmic ones when seeking advice for managing a health condition. Even though more

[6] We remain agnostic on whether or not blanket technological acceptance contributes to societal good or ill within the scope of this paper, but encourage readers to explore the body of work on technological ethics (Müller, 2020).

deliberative individuals still preferred a greater amount of AI advice than intuitive individuals in this decision domain, it is of practical interest to those seeking to encourage adoption of novel algorithmic technologies in healthcare settings to consider framing them as humanly as possible. This is in contrast to prior research on therapy chatbots, where users disclose more to a chatbot therapist when framed as non-human due to the freedom from judgment afforded by the format (Lucas et al., 2017). For theoreticians, the domain differences observed in this paper, and in the literature at large, merit further investigation.

We believe our results raise an important question about how human bias may interact with algorithmic bias to reproduce inequality. If more deliberative people are more prone to preferring algorithmic sources of input over human ones, then could they be more prone to believe biased algorithms? Although analytical thinking is correlated with traditionally-held markers of "success" such as understanding science, financial impulse control, and second language acquisition (Sheremeta, 2018; Pennycook et al., 2015; Shtulman & McCallum 2014; Jamieson 1992), as well as lower prejudice and religiosity (Franks & Scherr 2017; Karadöller et al., 2015), algorithms that have baked-in bias may undermine these deliberative processes as humans increasingly "outsource" cognitive tasks to artificial intelligence (Brynjolfsson & McAfee, 2011; Dastin, 2018). We hope future research will examine how cognitive styles may interact with algorithmic bias to amplify social stratification, despite our best intentions.

Limitations and Future Directions

We hope our work will inspire more individual-level difference research in understanding technology use and algorithmic aversion. Although cognitive style is understood to be a stable trait, measured as in this research and correlated with other attributes and behaviors, individuals deliberate or rely on their intuition depending on the task at hand. Future research could manipulate the amount of reflection that individuals engage in to understand whether advisor perceptions are formed prior to or during tasks that generate different amounts of reflection or intuition. Cognitive style is difficult to manipulate. We found failed attempts at manipulating cognitive style (Deppe et al., 2015; Baron, 2015) and only two documented instances of experimenters successfully manipulating cognitive style. Of the two instances, one resorted to injecting participants with testosterone (Nave et al., 2017) and the other that prompted recall of an event where reflection or intuition led to good or bad outcomes was deemed finicky and unreliable (D. Rand, personal communication, 2020; Shenhav, Greene, & Rand, 2012). Despite this difficulty, we remain hopeful that ingenious researchers will pursue this as a future direction for research.

Another limitation of our current work lies in the absence of high versus low stakes decision scenarios. Our situations were hypothetical and therefore low stakes in reality, although decisions about healthcare and hiring could be imagined as high stakes. Would individuals rely on advisors differently depending on the stakes? Would the manager in the opening vignette of this paper prefer to blame the dismissal of employees on the algorithmic advisor? When would a fired employee or his peers blame the AI tool versus the manager? We encourage future work, especially in field settings, to test the applicability of this effect in both low- and high-stakes decision situations.

A critic could also point to the artificial simplification of the advice seeking paradigm as a possible limitation of this work. In order to maximize participant comprehension of the study materials, we sought to present a simple and straightforward decision task. However, since past research has shown that confidence in the accuracy of one’s intuition depends on the difficulty of the cognitive task (Gill et al., 1998), future research should systematically vary the complexity of the decision-making task to see whether intuitive individuals would still rely on the human advisors in difficult scenarios as well as the simple ones presented here.

In all of the studies used in this paper, the algorithmic tool was consistently presented in a support role to a user who had decision authority. Given that AI can threaten to replace human decision makers (Gamez-Djokic & Waytz, 2020), how would individuals of varying cognitive styles respond to AI agents in positions of power or authority? Since objectivity also mediated the relationship between cognitive style and algorithmic aversion, and objectivity could vary depending on power relationships – either at the individual level or at the level of the companies that create and profit from the algorithms – future research should consider AI agents in roles beyond the support role of advisor.

This research focused on advice seeking preferences and not on advice adoption. Logg, Gino, & Minson (2020) found that stated preferences differ from actual behaviors when adopting algorithmic feedback. Future work should extend the current findings to see whether cognitive style indeed affects advice adoption as well as stated preferences.

Our research showed a strong difference across decision domain, with healthcare decisions prompting a preference for a human advisor, relative to three other decision domains. Future work could examine domain area scope conditions to see whether the relationship would hold for other subjective decision areas, such as choosing a partner for a date (Logg et al., 2019), assessing the humor of a joke (Yeomans et al., 2019), or choosing a stylish outfit (e.g., Stitchfix). Systematically contrasting decision domains across objective versus subjective, high versus low stakes, personal versus impersonal, socially visible versus private, and a host of other dimensions will clarify the boundary conditions in which cognitive style can predict algorithmic aversion by individuals.

Decision domains are not inherently subjective or objective but are construed by decision makers: Financial investment planning could be construed as a subjective decision about individual dreams and future scenarios or an objective decision about wealth maximization, and therefore experimental manipulation of construals could offer an approach to understanding algorithmic aversion. In addition, the broader level concept of ‘construal level’ may also be a factor that is linked to cognitive style, where individuals with a higher default construal level are also more intuitive, as they grasp the bigger picture through feel and instinct as opposed to a systematic consideration of the details (Kim & Duhachek, 2020).

Conclusion

Despite limitations, our studies contribute to the algorithmic decision-making literature in identifying a novel individual level difference – cognitive style – that influences algorithmic aversion. Our results add to the cognitive style literature in showing a behavioral difference between those who rely predominantly on their intuitive System 1 thinking and those who are more deliberative and rely on their System 2 thinking. This effect was driven by differential perceptions of advisor accuracy and, to a lesser extent, impartiality, where cognitively intuitive individuals believed human advisors to be more accurate and algorithmic advisors to be less impartial, whereas deliberative individuals believed the opposite. Algorithmic aversion was not predicted by attributes one might assume would be related, such as prior experience with AI, age, education, or gender, when cognitive style was being used as a predictor. Our results thus underscore the importance of not only focusing on the real accuracy of an algorithmic advisor, but of successfully communicating and contrasting the accuracy of the advisor against human equivalents, possibly through tailoring to the cognitive style of the decision-maker.

As one of the first papers on algorithmic aversion to focus on an individual level difference, we hope that this can open the door for more research to understand features of the user, to complement studies of features of the technology, that influence algorithmic aversion. We are just beginning to understand how cognitive style intersects with human reliance on technology, both as a predictor of use, and as an outcome (Vujic, 2017). This issue is of great importance, especially as we seek to understand how human cognitive bias may be amplified or attenuated by artificial intelligence, and which populations may be more vulnerable to any deleterious effects.

References

Adams, J. (1959). Advice seeking of mothers as a function of need for cognition. *Child Development*, 30(1), 171-176.

Ægisdóttir, S., White, M. J., Spengler, P. M., Maugherman, A. S., Anderson, L. A., Cook, R. S., Nichols, C. N., Lampropoulous, G. K., Walker, B. S., Cohen, G., & Rush, J. D. (2006). The meta-analysis of clinical judgment project: Fifty-six years of accumulated research on clinical versus statistical prediction. *The Counseling Psychologist*, *34*(3), 341-382.

Ares, G., Mawad, F., Giménez, A., & Maiche, A. (2014). Influence of rational and intuitive thinking styles on food choice: Preliminary evidence from an eye-tracking study with yogurt labels. *Food Quality and Preference*, 31, 28-37.

Armstrong, S., Allinson, C., & Hayes, J. (2002). Formal Mentoring Systems: An Examination of the Effects of Mentor/Protégé Cognitive Styles on the Mentoring Process. *Journal of Management Studies*, *39*(8), 1111–1137. https://doi.org/10.1111/1467-6486.00326

Aspect. (2016, October 26). Aspect Customer Experience Survey. https://www. aspect. com/globalassets/2016-aspect-consumer-experience-index-survey_index-results-final. pdf

Bago, B., Rand, D. G., & Pennycook, G. (2020) Fake news, fast and slow: Deliberation reduces belief in false (but not true) news headlines. *Journal of experimental psychology: general,* 149(8), 1608–1613.

Bahçekapili, Hasan & Yılmaz, Onurcan. (2017). The relation between different types of religiosity and analytic cognitive style. *Personality and Individual Differences*. 117. 267-272.

Bainbridge, H. (2008). The effect of presence on human-robot interaction. *RO-MAN 2008 - The 17th IEEE International Symposium on Robot and Human Interactive Communication*, 701–706. https://doi. org/10. 1109/ROMAN. 2008. 4600749

Baron, R. M. and Kenny, D. A. (1986) "The Moderator-Mediator Variable Distinction in Social Psychological Research – Conceptual, Strategic, and Statistical Considerations", Journal of Personality and Social Psychology, Vol. 51(6), pp. 1173–1182.

Barr, N., Pennycook, G., Stolz, J. A., & Fugelsang, J. A. (2015). The brain in your pocket: Evidence that Smartphones are used to supplant thinking. *Computers in Human Behavior*, 48, 473–480.

Bartlett, R., Stanton, R., Morse, A., Wallace, N. (2019). *Consumer-Lending Discrimination in the FinTech Era*. Unpublished manuscript, https://doi. org/10. 3386/w25943

Berinsky, A. J., Huber, G. A., and Lenz, G. S. (2012). Evaluating online labor markets for experimental research: Amazon. com's Mechanical Turk. *Political analysis*, *20*(3), 351-368.

Bialek, M., & Pennycook, G. (2018). The cognitive reflection test is robust to multiple exposures. *Behavior research methods*, *50*(5), 1953-1959.

Białek, M., & Sawicki, P. (2018). Cognitive reflection effects on time discounting. *Journal of Individual Differences*, 39(2), 99–106.

Blohm, I., Antretter, T., Wincent, J., Sirén, C., & Grichnik, D. 2020. It's a Peoples Game, Isn't It?! A Comparison between the Investment Returns of Business Angels and Machine Learning Algorithms. *Entrepreneurship Theory and Practice*. September. doi:10.1177/1042258720945206

Bornstein, R. F., & D'Agostino, P. R. (1992) Stimulus recognition and the mere exposure effect. *Journal of personality and social psychology*, 63. 4, 545.

Brynjolfsson, E., & McAfee, A. (2011) *Race against the machine: How the digital revolution is accelerating innovation, driving productivity, and irreversibly transforming employment and the economy*, Brynjolfsson and McAfee.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2016). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality data? In A. E. Kazdin (Ed.), *Methodological issues and strategies in clinical research* (p. 133–139). American Psychological Association.

Cacioppo, J., Petty, R., Feinstein, J., & Jarvis, W. (1996). Dispositional Differences in Cognitive Motivation: The Life and Times of Individuals Varying in Need for Cognition. *Psychological Bulletin*, 119(2), 197-253.

Campitelli, G., & Labollita, M. (2010). Correlations of cognitive reflection with judgments and choices. *Judgment and Decision-making*. 5. 182-191.

Cassano, J. (2019, April 16). *NYC students take aim at segregation by hacking an algorithm*. Fast Company, https://www.fastcompany.com/90331368/nyc-students-take-aim-at-segregation-by-hacking-an-algorithm

Dastin, J. (2018, October 10). Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS quarterly*, 319-340.

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science,* 243(4899), 1668-1674.

De Keersmaecker, J., Onraet, E., Lepouttre, N., & Roets, A. (2017). The opposite effects of actual and self-perceived intelligence on racial prejudice. *Personality and Individual Differences*, *112*, 136-138.

Dietvorst, B. J., & Bharti, S. (2020) People Reject Algorithms in Uncertain Decision Domains Because They Have Diminishing Sensitivity to Forecasting Error. *Psychological Science*, *31*(10),1302-1314.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2016) Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General, 144*(1), 114.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2018) Overcoming Algorithm Aversion: People Will Use Imperfect Algorithms If They Can (Even Slightly) Modify Them. *Management Science, 64*(3), 1155–1170.

Dweck, C. S. (1986). Motivational processes affecting learning. *American psychologist*, *41*(10), 1040.

Dweck, C. S., Chiu, C. -Y., and Hong, Y. -Y. (1995). Implicit Theories and Their Role in Judgments and Reactions: A Word From Two Perspectives. *Psychological Inquiry*, 6(4), 267–285. http://doi. org/10. 1207/s15327965pli0604_1

Efendić, E., van de Calseyde, P. P. F. M., & Evans, A. M. (2020). Slow response times undermine trust in algorithmic (but not human) predictions. *Organizational Behavior and Human Decision Processes*, 157, 103-114.

Elsbach, K. D., and Stigliani, I. (2019). New information technology and implicit bias. *Academy of Management Perspectives*, *33*(2), 185-206.

Evans, J. S. B., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on psychological science, 8*(3), 223-241.

Fergus, T. A., Valentiner, D. P., McGrath, P. B., Gier-Lonsway, S. L., & Kim, H. S. (2012). Short-forms of the Social Interaction Anxiety Scale and the Social Phobia Scale. *Journal of Personality Assessment, 94*, 310-320. doi:10. 1080/00223891. 2012. 660291

Fergus, T. A., Valentiner, D. P., Kim, H. -S., & McGrath, P. B. (2014). The Social Interaction Anxiety Scale (SIAS) and the Social Phobia Scale (SPS): A comparison of two short-form versions. *Psychological Assessment, 26*(4), 1281–1291. https://doi. org/10. 1037/a0037313

Franks, A., & Scherr, K. (2017). Analytic Thinking Reduces Anti-Atheist Bias in Voting Intentions. *The International Journal for the Psychology of Religion*. 10. 1080/10508619. 2017. 1313013.

Frederick, S. (2005) Cognitive reflection and decision-making. *Journal of Economic perspectives*, 19. 4, 25-42.

Fu, H., Huang, Y., Vir Singh, P. (2020). *Crowd, Lending, Machine, and Bias*. Unpublished Manuscript, http://dx. doi. org/10. 2139/ssrn. 3206027

Gill, M. J., Swann, W. B., Jr., & Silvera, D. H. (1998). On the genesis of confidence. *Journal of Personality and Social Psychology*, 75, 1101– 1114.

Glikson, E., & Woolley, A. (2020). Human Trust in Artificial Intelligence: Review of Empirical Research. *The Academy of Management Annals*, 14 (2).

Godek, J., and Murray, K. (2008). Willingness to pay for advice: The role of rational and experiential processing. *Organizational Behavior and Human Decision Processes*, 106(1), 77-87.

Golden, J. A. (2017). Deep learning algorithms for detection of lymph node metastases from breast cancer: helping artificial intelligence be seen. *Jama*, *318*(22), 2184-2186.

Hayes, A. F. (2012). PROCESS: A versatile computational tool for observed variable mediation, moderation, and conditional process modeling [White paper]. Retrieved from http://www. afhayes. com/public/process2012. pdf

Hayes, J., & Allinson, C. W. (1994). Cognitive style and its relevance for management practice. *British journal of management*, *5*(1), 53-71.

Heimberg, R. G., Horner, K. J., Juster, H. R., Safren, S. A., Brown, E. J., Schneier, F. R., & Liebowitz, M. R. (1999). Psychometric properties of the Liebowitz social anxiety scale. *Psychological medicine*, *29*(1), 199-212.

Hoppe, E. I., & Kusterer, D. J. (2011). Behavioral biases and cognitive reflection. Econ. Lett. 110, 97–100. doi: 10. 1016/j. econlet. 2010. 11. 015

Horton, J. J., Rand, D. G., & Zeckhauser, R. J. (2011). The online laboratory: Conducting experiments in a real labor market. *Experimental economics*, 14(3), 399-425.

Hutson, J. A., Taft, J. G., Barocas, S., & Levy, K. (2018). Debiasing desire: Addressing bias & discrimination on intimate platforms. *Proceedings of the ACM on Human-Computer Interaction, 2*(CSCW), 1-18.

Iyer, R., Koleva, S., Graham, J., Ditto, P., & Haidt, J. (2012). Understanding libertarian morality: The psychological dispositions of self-identified libertarians. *PLoS ONE*, 7(8), Article e42366.

Jacobson, H. K. (1969). Mass media believability: A study of receiver judgments. *Journalism quarterly*, *46*(1), 20-28.

Jamieson, J. (1992). The cognitive styles of reflection/impulsivity and field independence/dependence and ESL success. *The Modern Language Journal*, *76*(4), 491-501.

John, O. P., & Srivastava, S. (1999). The Big 5 trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin and O. P. John (Eds.), *Handbook of personality: Theory and Research* (2nd ed., pp. 102-138). New York: Guilford.

Kahneman, D. (2011) *Thinking, fast and slow*. Macmillan

Karadöller, D., Yılmaz, O., & Sofuoglu, G. (2015). Analytic Thinking, Religion and Prejudice: An Experimental Testing of the Dual-Process Model of Mind. *The International Journal for the Psychology of Religion*, 26(4), 360–369

Kaufmann, E., Budescu, D. V. (2020). Do Teachers Consider Advice? On the Acceptance of Computerized Expert Models. *Journal of Educational Measurement, 57*(2), 311–342.

Kim, C. (2016). Anthropomorphized Helpers Undermine Autonomy and Enjoyment in Computer Games. *The Journal of Consumer Research*, 43(2), 282–302. https://doi.org/10.1093/jcr/ucw016

Kim, S. (2019). Eliza in the uncanny valley: anthropomorphizing consumer robots increases their perceived warmth but decreases liking. *Marketing Letters*, 30(1), 1–12. https://doi.org/10.1007/s11002-019-09485-9

Kim, T. W., & Duhachek, A. (2020). Artificial Intelligence and Persuasion: A Construal-Level   Account. *Psychological Science*, 31(4), 363-380.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology, 77*(6), 1121–1134. https://doi.org/10.1037/0022-3514.77.6.1121

Kuncel, N. R., Klieger, D. M., Connelly, B. S., & Ones, D. S. (2013). Mechanical versus clinical data combination in selection and admissions decisions: A meta-analysis. *Journal of applied psychology*, *98*(6), 1060.

Lee, Y., Kozar, K. A., & Larsen, K. R. (2003). The technology acceptance model: Past, present, and future. *Communications of the Association for information systems*, *12*(1), 50.

Li, D., Raymond, L. R., & Bergman, P. (2020). *Hiring as Exploration* (No. w27736). National Bureau of Economic Research.

Li, R. (2010). A Cross-cultural Study: Effect of Robot Appearance and Task. *International Journal of Social Robotics*, 2(2), 175–186. https://doi. org/10. 1007/s12369-010-0056-9

Liebowitz, M. R. (1987). Social Phobia, Modern Problems of Pharmacopsychiatry, 22, 141-173.

Liu, Y., Kohlberger, T., Norouzi, M., Dahl, G. E., Smith, J. L., Mohtashamian, A., Olson, N., Peng, L. H., Hipp, J. D., & Stumpe, M. C. (2019). Artificial intelligence–based breast cancer nodal metastasis detection: Insights into the black box for pathologists. *Archives of pathology & laboratory medicine*, *143*(7), 859-868.

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, 151, 90-103.

Logg, J. M., Gino, F., & Minson, J. A. (2020). Robo-Coaching: When do people prefer performance assessments from algorithms versus people? Manuscript in preparation.

Longoni, C., Bonezzi, A., & Morewedge, C. K. (2018). Consumer Reluctance Toward Medical Artificial Intelligence: the Underlying Role of Uniqueness Neglect. *ACR North American Advances*.

Lucas, G. M., Rizzo, A., Gratch, J., Scherer, S., Stratou, G., Boberg, J., & Morency, L. P. (2017). Reporting mental health symptoms: breaking down barriers to care with virtual human interviewers. *Frontiers in Robotics and AI*, *4*, 51.

Maier, S. R. (2005). Accuracy matters: A cross-market assessment of newspaper error and credibility. *Journalism & Mass Communication Quarterly*, *82*(3), 533-551.

Mawad, F., Trías, M., Giménez, A., Maiche, A., Ares, G. (2015). Influence of cognitive style on information processing and selection of yogurt labels: Insights from an eye-tracking study. *Food Research International,* 74, 1–9.

Marr, 2018. https://www. forbes. com/sites/bernardmarr/2018/07/25/how-is-ai-used-in-education-real-world-examples-of-today-and-a-peek-into-the-future/#146ee5a8586e

McCrae, R. R., & Costa, P. T. (1987). Validation of the five-factor model of personality across instruments and observers. *Journal of personality and social psychology*, *52*(1), 81.

Metawa, H., Hassan, M. K., & Elhoseny, M. (2017). Genetic algorithm based model for optimizing bank lending decisions. *Expert Systems with Applications*, 80, 75–82. https://doi. org/10. 1016/j. eswa. 2017. 03. 021

Meyer, A., Zhou, E., & Frederick, S. (2018). The non-effects of repeated exposure to the Cognitive Reflection Test. *Judgment and Decision-making*, *13*(3), 246.

Müller, V. C. (2020). Ethics of Artificial Intelligence and Robotics. In E. A. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*, (Winter edition), https://plato. stanford. edu/archives/win2020/entries/ethics-ai/

Nadler, A., Jiao, P., Johnson, C., Alexander, V., Zak, P. (2018). The Bull of Wall Street: Experimental Analysis of Testosterone and Asset Trading. *Management Science*, 64, 4032-4051.

Neu, W. A., Gonzalez, G. R., & Pass, M. W. (2011). The trusted advisor in inter-firm interpersonal relationships. *Journal of Relationship Marketing*, *10*(4), 238-263.

Noble, S. (2018) *Algorithms of Oppression: How search engines reinforce racism*. NYU Press.

O'Neil, C. (2016) *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books.

Obermeyer, Z., Powers, B., Vogeli, C., Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations.

Önkal, D., Gönül, M. S., De Baets, S. (2019) Trusting forecasts. *Futures & Foresight Science*, 1(3-4)

Pak, F. (2012). Decision support aids with anthropomorphic characteristics influence trust and performance in younger and older adults. *Ergonomics*, 55(9), 1059–1072. https://doi. org/10. 1080/00140139. 2012. 691554

Pennycook, G., Cheyne, J. A., Seli, P., Koehler, D. J., & Fugelsang, J. A. (2012). Analytic cognitive style predicts religious and paranormal belief. *Cognition*, *123*(3), 335-346.

Pennycook, G., Cheyne, J. A., Barr, N., Koehler, D. J., & Fugelsang, J. A. (2015). On the reception and detection of pseudo-profound bullshit. *Judgment and Decision-making, 10*(6), 549-563.

Pennycook, G., & Rand, D. G. (2020) Who falls for fake news? The roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking. *Journal of personality,* 88. 2, 185-200.

Pennycook, G., & Rand, D. G. (2019) Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, 188, 39-50.

Pennycook, G., Fugelsang, J. A., & Koehler, D. J. (2015). Everyday consequences of analytic thinking. *Current Directions in Psychological Science*, 24(6), 425–432.

Pornpitakpan, C. (2004). The persuasiveness of source credibility: A critical review of five decades' evidence. *Journal of applied social psychology*, *34*(2), 243-281.

Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Vaughan, J. W., & Wallach, H. (2018). Manipulating and measuring model interpretability. *arXiv preprint arXiv:1802.07810*.

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

Rand, D. G. (2012). The promise of Mechanical Turk: How online labor markets can help theorists run behavioral experiments. *Journal of theoretical biology*, *299*, 172-179.

Rammstedt, B., & John, O. P. (2007). Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of research in Personality, 41*(1), 203-212.

Rattan, A., & Georgeac, O. A. M. (2017). Understanding intergroup relations through the lens of implicit theories (mindsets) of malleability. *Social and Personality Psychology Compass,* 11(4), Article e12305. https://doi. org/10. 1111/spc3. 12305

Riva, P., Sacchi, S., & Brambilla, M. (2015). Humanizing machines: Anthropomorphization of slot machines increases gambling. *Journal of Experimental Psychology: Applied*, *21*(4), 313.

Rodriguez, M. C., Ooms, A., & Montañez, M. (2008). Students’ perceptions of online-learning quality given comfort, motivation, satisfaction, and experience. *Journal of interactive online learning*, *7*(2), 105-125.

Rodriguez-Ruiz, A., Lång, K., Gubern-Merida, A., Broeders, M., Gennaro, G., Clauser, P., Helbich, T. H., Chevalier, M., Tan, T., Mertelmeier, T., Wallis, M. G., Andersson, I., Zackrisson, S., Mann, R. M., Sechopoulos, I. (2019). Stand-alone artificial intelligence for breast cancer detection in mammography: comparison with 101 radiologists. *JNCI: Journal of the National Cancer Institute*, *111*(9), 916-922.

Ross, R. M., Pennycook, G., McKay, R., Gervais, W. M., Langdon, R., & Coltheart, M. (2016). Analytic cognitive style, not delusional ideation, predicts data gathering in a large beads task study. *Cognitive Neuropsychiatry*, *21*(4), 300-314.

Rucker, D. D., Preacher, K. J., Tormala, Z. L., & Petty, R. E. (2011). Mediation analysis in social psychology: Current practices and new recommendations. *Social and Personality Psychology Compass*, *5*(6), 359-371.

Shenhav, A., Rand, D. G., Greene, J. D. (2012). Divine Intuition: Cognitive Style Influences Belief in God. *Journal of Experimental Psychology*. *General*, 141(3), 423–428.

Sheremeta, R. M. (2018) Impulsive Behavior in Competition: Testing Theories of Overbidding in Rent-Seeking Contests, Available at SSRN: https://ssrn. com/abstract=2676419 or http://dx. doi. org/10. 2139/ssrn. 2676419

Shtulman, A., & McCallum, K. (2014). Cognitive reflection predicts science understanding. *Proceedings of the 36th Annual Conference of the Cognitive Science Society*, 2937–2942.

Sinayev A., & Peters, E . (2015). Cognitive reflection vs. calculation in decision making. Front. Psychol. 6:532. doi: 10. 3389/fpsyg. 2015. 00532

Simonite, T. (2020, August 10). *Meet the Secret Algorithm That's Keeping Students Out of College*. Wired, https://www. wired. com/story/algorithm-set-students-grades-altered-futures/

Steen, L. A. (1990). Numeracy. *Daedalus*, 211-231.

Stupple, E. J. N., Pitchford, M., Ball, L. J., Hunt, T. E., Steel, R. (2017) Slower is not always better: Response-time evidence clarifies the limited role of miserly information processing in the Cognitive Reflection Test. PloS One, 12(11).

Thoma, V., White, E., Panigrahi, A., Strowger, V., & Anderson, I. (2015). Good thinking or gut feeling? Cognitive reflection and intuition in traders, bankers and financial non-experts. *PLoS ONE*, 10(4), Article e0123202.

Thomson, K. S., & Oppenheimer, D. M. (2016). Investigating an alternate form of the cognitive reflection test. *Judgment and Decision Making*, 11(1), 99–113.

Toplak, M. E., West, R. F., & Stanovich, K. E. (2014) Assessing miserly information processing: An expansion of the Cognitive Reflection Test. " *Thinking and Reasoning*, 20. 2, 147-168.

Toplak, M. E., West, R. F., & Stanovich, K. E. (2011). The Cognitive Reflection Test as a predictor of performance on heuristics-and-biases tasks. *Memory and cognition*, *39*(7), 1275.

Turner Lee, N. (2018). Detecting racial bias in algorithms and machine learning. *Journal of Information, Communication & Ethics in Society* (Online), 16(3), 252–260. https://doi. org/10. 1108/JICES-06-2018-0056

Waytz, A., Heafner, J., & Epley, N. (2014). The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology*, *52*, 113-117.

Vander Ark, T. (2018, August 10). *32 Ways AI is Improving Education*. Getting Smart, https://www. gettingsmart. com/2018/08/32-ways-ai-is-improving-education/

Vujic, A. (2017). Switching on or switching off? Everyday computer use as a predictor of sustained attention and cognitive reflection. *Computers in Human Behavior*, 72, 152-162.

Weinstock, O. (2012). The effect of system aesthetics on trust, cooperation, satisfaction and annoyance in an imperfect automated system. *Work (Reading, Mass.)*, 41 Suppl 1, 258–265. https://doi. org/10. 3233/WOR-2012-0166-258

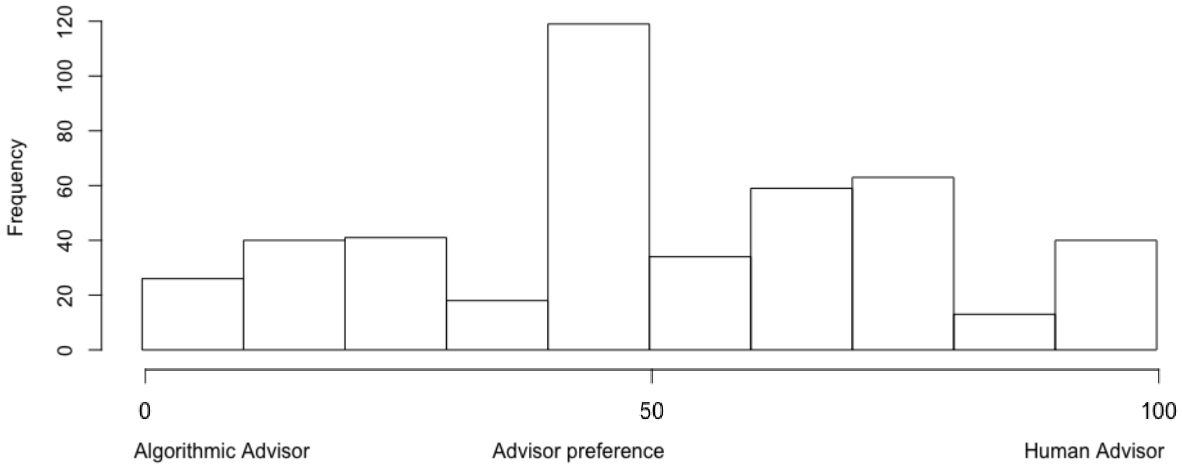
Yeomans, M., Shah, A., Mullainathan, S., & Kleinberg, J. (2019). Making sense of recommendations. *Journal of Behavioral Decision-making.*

Yılmaz, O., & Sarıbay, S. A. (2018). The relationship between cognitive style and political orientation depends on the measures used. *Social Psychology*, 49(2), 65–75 https://doi. org/10. 1027/1864-9335/a000328

Zhang, B., & Dafoe, A. (2019). Artificial intelligence: American attitudes and trends. *Available at SSRN 3312874*.

Figures

Figure 1. Distribution of choice frequency for preferred advice proportion



Note. Answers of 0-10 are combined in the first bar.

Figure 2. Distribution of frequency counts for the Standardized CRT

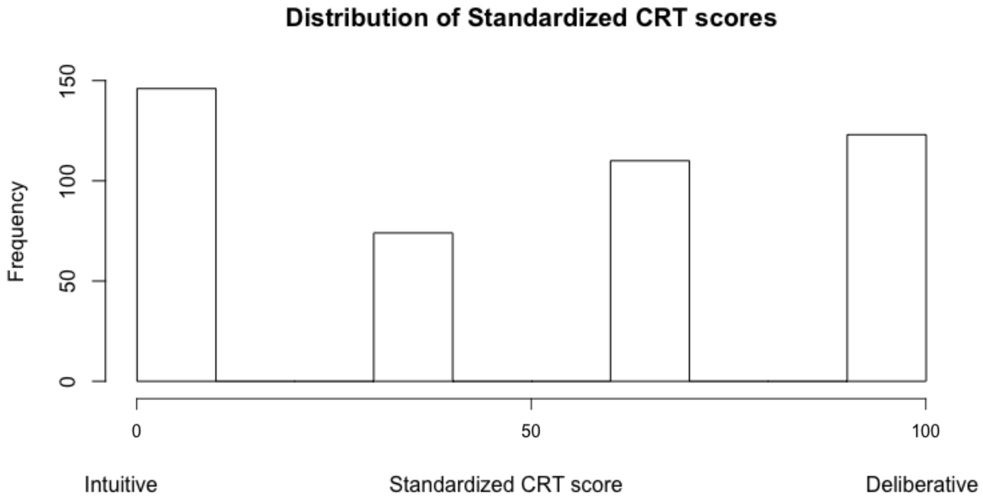
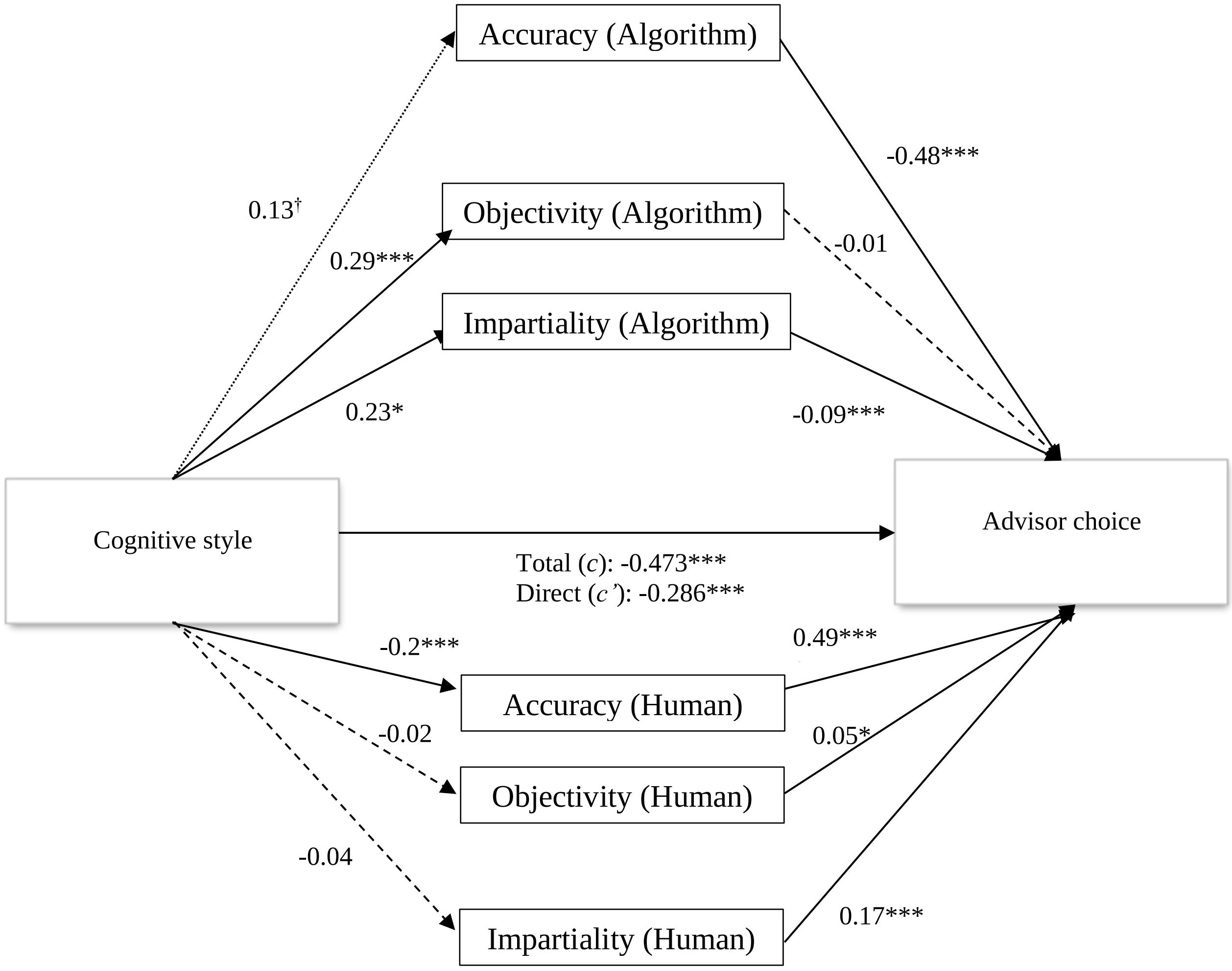


Figure 3. Multiple mediation model showing relationship between cognitive style and algorithmic aversion in advisor choice.



Tables

Table 1. Correlation Table of Variables Associated with Algorithmic Aversion for Study 1

	Algorith-aversion	CRT score	Comfort with tech.	Social anxiety	Extraversion	Conscientiousness	Agreeableness	Neuroticism	Openness	Confidence in fin. liter.	Age	Income
CRT score	-0.21****											
Comfo. with tech	-0.17***	0.18***										
Social anxiety	0.09*	-0.17***	-0.19****									
Extraversion	0.09	-0.02	-0.01	-0.44****								
Conscient.	0.08	0.06	0.16***	-0.31****	0.17***							
Agreeableness	0.06	0.05	0.09	-0.24****	0.09*	0.21****						
Neuroticism	0.01	-0.11*	-0.10*	0.41****	-0.37****	-0.33****	-0.26****					

	Algorith-aversion	CRT score	Comfort with tech.	Social anxiety	Extraversion	Conscientiousness	Agreeableness	Neuroticism	Openness	Confidence in fin. liter.	Age	Income
Openness	-0.02	0.15**	0.19****	-0.21****	0.11*	0.15**	0.13**	-0.07				
Confidence in fin. literacy	-0.06	0.09	0.24****	-0.16***	0.14**	0.21****	0.14**	-0.32****	0.01			
Age	-0.04	0.09*	-0.11*	-0.29****	0.05	0.21****	0.13**	-0.20****	-0.03	0.01		
Income	-0.08	0.10*	0.06	-0.06	0.08	0.16***	0.04	-0.19****	-0.12*	0.14**	0.07	
Edu.	-0.06	0.19****	0.06	-0.04	0.03	0.04	0.07	-0.09*	0.10*	0.08	0.06	0.26 ****

* p < 0.05, ** p < 0.01, *** p < 0.001

Table 2. OLS Regressions Predicting Relative Preference for Advice from Human (vs. Algorithmic) Advisor, controlling for individual differences for Study 1

	Model 1a	Model 1b	Model 1c	Model 1d
CRT Score	-13.700*** (2.946)	-11.047*** (2.996)	-12.003*** (3.205)	-9.361*** (3.233)
Comfort with Technology		-0.590** (0.216)		-0.644** (0.223)
Social Anxiety		0.649** (0.247)		0.605* (0.258)
Confidence in Financial Literacy		-1.027 (1.291)		-1.538 (1.367)
BIG-5				
Agreeableness		1.134† (0.655)		1.053 (0.666)
Conscientiousness		1.927** (0.720)		2.175** (0.768)
Extraversion		1.463* (0.600)		1.204† (0.618)
Neuroticism		0.254 (0.606)		-0.226 (0.647)
Openness		0.227 (0.606)		0.115 (0.656)
Age			-0.187 (0.130)	-0.237† (0.136)
Race *(Reference Group: White)*				
Black			1.581 (4.112)	3.043 (4.038)
American Indian			11.459 (11.762)	11.615 (11.597)
Asian			-3.381 (4.691)	-2.530 (4.665)
Other			-5.781 (7.233)	-5.931 (7.137)
Gender *(Reference Group: Male)*				
Female			2.835 (2.497)	2.337 (2.553)
Other Gender			22.254 (18.127)	15.342 (18.051)
Employment Status *(Reference Group: Employed)*				
Unemp. Search			5.905 (4.800)	7.782 (4.761)
Unemp. Not search			-5.056 (5.608)	-5.785 (5.608)

	Model 1a	**Model 1b**	**Model 1c**	**Model 1d**
Retired			12.637†	12.203
			(7.551)	(7.512)
Student			-7.672	-7.680
			(6.237)	(6.139)
Political Affiliation *(Reference Group: Republican)*				
Democrat			-1.596	-1.753
			(3.201)	(3.243)
Independent			-7.305*	-7.448*
			(3.553)	(3.525)
Other Party			-14.142	-17.127
			(13.578)	(13.404)
None			-9.830	-9.218
			(7.829)	(7.768)
Income			0.000	0.000
			(0.000)	(0.000)
Education *(Reference Group: No High School)*				
HS			-16.151	-17.180
			(13.678)	(13.540)
Some College			-16.250	-15.523
			(13.368)	(13.327)
AA			-16.980	-18.362
			(13.618)	(13.565)
BA			-16.315	-16.859
			(13.390)	(13.290)
MA			-18.338	-18.207
			(13.593)	(13.460)
PhD			-16.854	-12.379
			(20.296)	(20.176)
JD/MD			-14.786	-16.216
			(15.657)	(15.545)
Constant	62.357***	33.308*	87.752***	68.896***
	(1.858)	(13.349)	(14.494)	(20.715)
R^2	0.046	0.102	0.092	0.151
Adjusted R^2	0.044	0.084	0.041	0.086
F Statistic	21.629	5.581	1.811	2.327

† $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note. This table shows ordinary least squares (OLS) regressions predicting relative proportion of advice from a human (vs. algorithmic) advisor using an indicator of individual cognitive reflection. Negative coefficients denote preference for greater advice from algorithmic advisor. Standard errors are in parentheses.

Table 3. Correlation Table of Variables Associated with Algorithmic Aversion for Study 2

	Algorithmic aversion	**CRT score**	**Social anxiety**	**Intelligence**	**Mindset**	**Age**	**Income**
CRT score	-0.23**						
Social anxiety	0.07	-0.06					
Intelligence	0.15*	0.04	0.02				
Mindset	0.12	-0.02	0.21**	0.20**			
Age	0.04	-0.12	-0.19**	0.07	0.01		
Income	0.11	0.04	-0.09	0.08	0.02	-0.05	
Education	0.08	0.11	0.03	0.22**	0.17*	0.07	0.31***

* p < 0.05, ** p < 0.01, *** p < 0.001

Table 4. OLS Regressions Predicting Relative Preference for Advice from Human (vs. Algorithmic) Advisor, controlling for individual differences for Study 2

	Model 2a	Model 2b	Model 2c
CRT Score	-19.125**	-19.248***	-14.149*
	(5.767)	(5.725)	(6.351)
Social Anxiety		0.062	0.085
		(0.118)	(0.122)
Self-perceived Intelligence		0.239*	0.172
		(0.120)	(0.131)
Growth Mindset		0.516	0.228
		(0.434)	(0.474)
Constant	65.291***	44.746***	48.271*
	(3.324)	(9.119)	(22.629)
Demographic Controls	N	N	Y
R^2	0.051	0.084	0.203
Adjusted R^2	0.046	0.066	0.088
F Statistic	10.999	4.640	1.764

† $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note. This table shows ordinary least squares (OLS) regressions predicting relative proportion of advice from a human (vs. algorithmic) advisor using an indicator of individual cognitive reflection. Negative coefficients denote preference for greater advice from algorithmic advisor. Standard errors are in parentheses.

Table 5. Correlation Table of Variables Associated with Algorithmic Aversion for Study 3

	Algorithmic aversion	**CRT score**	**Intelligence**	**Prior exp. w/AI**	**Age**	**Income**
CRT score	-0.14*					
Intelligence	-0.09	0.15**				
Prior exp. w/AI	0.01	0.02	0.05			
Age	-0.04	0.08	-0.03	0.12*		
Income	0.04	0.10	0.15**	0.00	0.04	
Education	0.00	0.15**	0.15**	-0.01	0.04	0.40****

* p < 0.05, ** p < 0.01, *** p < 0.001

Table 6. OLS Regressions Predicting Relative Preference for Advice from Human (vs. Algorithmic) Advisor, controlling for individual differences for Study 3

	Model 3a	Model 3b	Model 3c	Model 3d	Model 3e
CRT Score	-10.691*** (2.808)	-9.513*** (2.807)	-9.040** (2.908)	-14.675** (5.442)	-15.406** (5.542)
Domain					
College Admissions		4.393 (3.061)	5.067 (3.114)	1.671 (4.798)	1.052 (4.884)
Healthcare Management		12.108*** (3.026)	11.821*** (3.054)	9.739* (4.500)	8.819† (4.516)
Hiring		4.630 (3.009)	4.569 (3.008)	0.498 (4.573)	0.192 (4.563)
Self-perceived Intelligence		-0.110† (0.066)	-0.109 (0.069)	-0.110† (0.066)	-0.108 (0.069)
CRT x College Admissions Domain				6.101 (8.068)	8.963 (8.211)
CRT x Healthcare Domain				5.378 (7.842)	6.928 (7.917)
CRT x Hiring Domain				9.090 (7.541)	9.722 (7.582)
Constant	67.061*** (1.659)	68.359*** (4.819)	69.189*** (7.627)	70.672*** (5.233)	71.736*** (7.878)
Demographic Controls	N	N	Y	N	Y
R^2	0.028	0.063	0.124	0.066	0.128
Adjusted R^2	0.026	0.054	0.073	0.051	0.071
F Statistic	14.500	6.798	2.446	4.423	2.268

† $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note. This table shows ordinary least squares (OLS) regressions predicting relative proportion of advice from a human (vs. algorithmic) advisor using an indicator of individual cognitive reflection. Negative coefficients denote preference for greater advice from algorithmic advisor. Standard errors are in parentheses.

Table 7. Correlation table of Advisor Perception Traits for Study 4

	Accuracy – Human	**Impartiality – Algo.**	**Impartiality – Human**	**Objectivity – Alg.**	**Objectivity – Human**
Accuracy – Human	0.19****				
Impartiality – Algo.	0.36****	0.05			
Impartiality – Human	-0.04	0.38****	-0.03		
Objectivity – Alg.	0.22****	0.11***	0.21****	0.09**	
Objectivity – Human	0.06*	0.29****	0.02	0.30****	-0.19****

* p < 0.05, ** p < 0.01, *** p < 0.001

Table 8. OLS Regressions Predicting Relative Preference for Advice from Human (vs. Algorithmic) Advisor, controlling for Advisor Perceptions for Study 4

	Model 4a	**Model 4b**	**Model 4c**	**Model 4d**	**Model 4e**
CRT Score	-11.738***	-11.71***	-12.288***	-7.103***	-6.090***
	(1.931)	(1.93)	(2.501)	(1.509)	(1.571)
Order of CRT		2.702†	2.128		
		(1.534)	(2.203)		
CRT x Order			1.428		
			(3.933)		
Prior Experience with AI				0.025	0.062
				(0.606)	(0.607)
Perceived Advisor Accuracy					
Algorithmic				-11.847***	-11.875***
				(0.721)	(0.728)
Human				12.130***	11.835***
				(0.876)	(0.890)
Perceived Advisor Impartiality					
Algorithmic				-2.118***	-2.159***
				(0.561)	(0.564)
Human				4.110***	4.118***
				(0.636)	(0.636)
Perceived Advisor Objectivity					
Algorithmic				-0.179	-0.026
				(0.569)	(0.575)
Human				1.228*	1.594**
				(0.592)	(0.595)
Constant	62.120***	61.030***	61.264***	50.231***	44.970***
	(1.082)	(1.246)	(1.403)	(4.906)	(7.955)
Demographic Controls	N	N	N	N	Y
N	1052	1052	1052	1052	1052
R^2	0.034	0.037	0.037	0.426	0.451
Adjusted R^2	0.033	0.035	0.034	0.423	0.362
F Statistic	36.940	20.058	13.405	155.116	5.099

† $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note. This table shows ordinary least squares (OLS) regressions predicting relative proportion of advice from a human (vs. algorithmic) advisor using an indicator of individual cognitive reflection. Negative coefficients denote preference for greater advice from algorithmic advisor. Standard errors are in parentheses.

Table 9. Correlation Table for Study 4

	Alg. aversion	CRT score	Quality exp. w/ AI	Accuracy – Alg.	Accuracy – Human	Impart. – Alg.	Impart. – Hum.	Obj. – Alg.	Obj. – Hu.	Age	Income
Algorithmic aversion											
CRT score	-0.18****										
Quality exp. w/ AI	-0.15****	0.03									
Accuracy – Alg.	-0.41****	0.05	0.20****								
Accuracy – Human	0.37****	-0.10***	-0.07*	0.19****							
Impartiality – Alg.	0.35****	-0.02	-0.10**	-0.04	0.38****						
Impartiality – Human	-0.25****	0.08*	0.16****	0.36****	0.05	-0.03					
Objectivity – Alg.	-0.09**	0.10***	0	0.22****	0.11***	0.09**	0.21****				
Objectivity – Human	0.18****	-0.01	0	0.06*	0.29****	0.30****	0.02	-0.19****			
Age	-0.04	0.07*	-0.01	-0.04	-0.02	0	0.06	0.06	0.04		
Income	-0.02	0.11***	0.01	0.04	-0.01	-0.01	0.01	0.07*	-0.05	0.06*	
Education	-0.08**	0.16****	-0.01	-0.01	-0.07*	-0.02	-0.04	0.06*	0	0.10**	0.29 ****

* p < 0.05, ** p < 0.01, *** p < 0.001

CHAPTER 2: DIGITAL VOICED ASSISTANTS

The Facilitatory Effects of Stereotype-Congruent Features of Digital Voiced Agents

Heather Yang

Abstract

The use of female voices in novel technologies has been justified through a generalized preference for female voices. I draw upon the stereotype congruence literature to establish that, while individuals may explicitly justify their preference due to superficial characteristics of a female voice, their preferences are often in fact driven by stereotype congruence. Gender cues, as conveyed through a device's voice, can either be congruent or incongruent with the stereotypes associated with the role that the device is designed to serve, such as the female-typed role of assistant. I propose that congruent pairings serve to facilitate understanding of novel technologies by borrowing from categorical information of the stereotype, such that novice users better understand a device's capabilities. In this chapter, I investigate the facilitatory nature of stereotype-congruent features of digital voiced agents (DVAs), with a special focus on the match between the device's voice gender and the gender-typing of the device's job role. I discuss future directions for research and practical implications for user-interface designers and policy makers for artificially intelligent personified technology.

Introduction

With the rise of artificial intelligence and robotics, myriad anthropomorphized algorithmic agents have become embedded into everyday life. The most visible form of this development has been through the wide adoption of conversational digital voiced agents (DVAs), such as Apple's Siri, Amazon's Alexa, Google Assistant, and Microsoft's Cortana. In 2017, Pew American Trends reported that nearly half of all Americans (46%) use a DVA to interact with their smartphone or smart devices. Estimates for 2019 suggested that 3.25 billion digital voice assistants were being used worldwide (Vailshery, 2021a), with forecasts of over 8 billion units being used in 2023 (Vailshery, 2021b). Capitalizing on the large user base, the global intelligent virtual assistant market was valued between $2.2-$3.7 billion in 2019 (Grand View Research, 2020), with an upward growth trajectory predicted due to expansion into both commercial and industrial markets (Adroit Market Research, 2020).

Not only have DVAs found widespread use in households, but they have also gained interest from the commercial sector, from assisting in corporate scheduling (like x.ai), to monitoring the health of backend website infrastructure (through Amazon's Alexa). As their role in the workplace increases, so does the importance of understanding how these digital personalities affect, and are affected by, those who work with them.

The striking commonality among these bodiless agents is the default feminine characteristics assigned to them: all are female-voiced and female-named (with the exception of female-voiced but generically named "Google Assistant"). Even with the addition of male voice options in 2018 (Bonnington, 2018), the percentage of users who go through their settings to deliberately change the gender of their digital assistant from the default female voice is not collected or publicly available information (B. Auxier of the Pew Research Center, personal communication, April 2, 2020). In collecting base rates on American adult DVA users, I found that only 5% (9/170) of the users when asked reported that they changed the gender of the voice from the default female to male. This work seeks to investigate the reasons for the sustained use of female defaults in DVAs.

In this chapter, I empirically investigate the currently untested claim of whether gender stereotypes inform the preference for female voices for DVAs and answer the question of why individuals prefer women's voices for the most popular form of digital AI agents. In doing so, I present evidence of gender stereotype congruence as driving the preference for women's voices for female gender-typed roles (Studies 1 and 2) and examine whether theses stereotype-congruent personalities serve to facilitate the understanding of novel devices (Study 3), thus enabling their adoption. Before describing the studies in greater detail, I review the prior literature on stereotype congruence and the use of female voices in technologies to situate the current work.

Theoretical background

> *"For our objectives—building a helpful, supportive, trustworthy assistant—a female voice was the stronger choice."*
>
> - Microsoft spokeswoman speaking about Cortana to the WSJ, 2017.

> *"It's much easier to find a female voice that everyone likes than a male voice that everyone likes. It's a well-established phenomenon that the human brain is developed to like female voices."*

- Clifford Nass, Stanford University Professor, in an interview with CNN, 2011.

The Generalized Preference for Women's Voices

Technology companies and human-computer interaction researchers justify the use of female-defaults in DVAs using antiquated research that states that humans have an innate preference for female voices, starting in the womb. Clifford Nass, renowned for his pioneering work on Human-Computer Interaction, attributed the popularity of female voices embedded in technologies to their generalized preferability, as compared to either male or gender-neutral alternatives (Nass & Yen, 2010). In citing studies from evolutionary biology as well as his own laboratory work on the evaluation of female (vs. male) robots, he mirrors the widely accepted justification for the prevalence of the female defaults in digital voiced technologies (Nass, Moon, & Green, 1997). Work in developmental psychology compares the rate of responding of fetuses and newborns to their parents' voices and finds that newborns preferentially respond to the mother's, but not the father's, voice (Lee & Kisilevsky, 2013). Brain-imaging studies have identified different regions becoming activated when individuals listen to male vs. female voices, with women's voices triggering the auditory cortex more intensely than men's voices, leading researchers to believe that women's voices are easier to decode and listen to (Sokhi et al., 2005). Similarly, laboratory work on machine-synthesized voices finds that both male and female participants rate female voices as warmer and thus more preferable than male voices (Mitchell et al., 2011).

This academic literature also is consistent with the justifications for the female personalities of the most commonly available DVAs by their creators. Focus groups led by Amazon's Alexa team (Personal communication, 2017; Stern, 2017) and the developers of

Microsoft's Cortana (Stern, 2017) found that participants preferred the female voices over the male options, citing that women's voices were warmer and more approachable. The industry justification is consistent with the current academic literature in presenting a generalized preference for women's voices as driving the female default in DVAs.

Context Determines Preference for Women's Voices

However, this account – that women's voices are generally preferable – disregards the *role* in which the voice is being employed. In fact, female cues – like a female voice – can actually be less preferable compared to male cues in many instances. A robust literature has documented the preferability of male voices across a wide range of roles, usually male-typed and sometimes neutral roles, but mostly not female-typed roles. This literature also shows how, for these male-typed roles, identical content is penalized when it is associated with female cues. This pattern occurs in situations as far ranging as expert witnesses, blog credibility, political elections, and code pull requests on Github (Neal, Guadagno, Eno, & Brodsky, 2012; Yang et al., 2013; Armstrong & McAdams, 2009; Terrell et al., 2007). In the last example, programming code assigned female author names in Github had fewer pull requests accepted than when the same code was submitted anonymously. Klofstad, Anderson, and Nowicki (2015) found that male politicians were more likely to be elected than females of the same age in a hypothetical vote choice simulation with male and female voices, with perceived competence ratings mediating the effect. In their canonical paper showing increased speech speed being linked with perceptions of greater intelligence, Brown, Strong, and Rencher (1973) also found that having a lower-pitched voice (fundamental frequency) was linked with perceptions of increased competence. On the whole, extant justifications for the use of women's voices for these digital technologies insufficiently consider how female gender cues are interpreted across a range of roles and contexts.

Female cues are helpful, however, when the context calls for them. More facially-feminine women are seen as better fits for more feminine job titles (such as secretary or nurse; Johnson, Podratz, Dipboye, & Gibbons, 2010), and women are more likely to be promoted than men in feminine industries such as clothing manufacturing than in masculine industries such as auto manufacturing (Garcia-Retamero & Lopez-Zafra, 2006). In a field study on peer-to-peer lending, feminine-faced women received larger loans for domestic projects (like remodeling a kitchen) compared to those asking for loans to start or help run a business, or compared to their more masculine female peers (Kuwabara & Thebaud, 2017). So, although female cues can be favorable in certain contexts, such as for a feminine job role of a voiced assistant, it is likely that female cues would not be preferred for other contexts that are male-typed.

Stereotype Congruence

This pattern of rewarding roles or behaviors that are consistent with stereotypes around social identity characteristics (such as age, race, or gender) constitutes the phenomenon of ‘stereotype congruence’. Social cognition research has documented preferences for those who fall in line with role expectations, as opposed to defying them (Cialdini & Trost, 1998; Eagly & Karau, 2002). Stereotype-disconfirming information can cause personal threat to an individual’s worldview (Förster, Higgins, and Strack, 2000). Even if group members are negatively stereotyped, individuals prefer those who fulfill negative stereotypes over those who display positive characteristics that contrast with expectations for that group (Phelan & Rudman, 2010; Stern, West, and Rule, 2015).

A specific kind of stereotype congruence in the gender discrimination literature is Role Congruity Theory (Early & Karau, 2002), where members of a stereotyped group will be evaluated more positively when behaving or occupying roles that are in line with the social roles associated with their group (Rudman & Glick, 2001; Rudman et al., 2012). Considerable research has established that individuals preferentially match individuals to roles according to gender stereotypes (Heilman & Okimoto, 2007). This could be, for example, men in roles that have traditional masculine traits (e.g., agentic, assertive; Rudman et al., 2012). According to this theory, we expect that people would prefer a female voice for the feminine role of digital assistant due to the feminine associations with the role of administrative or personal assistant (Glick et al., 2005), rather than an innate preference for women's voices, as posed by the current industry justifications. This theory would suggest that the generalized preferability of women's voices is an incomplete explanation that omits the effect of stereotype congruence. Thus, the first hypothesis is:

H1: *Individuals will show gender stereotype congruence by preferring the female voice in female-typed roles, and the male voice in male-typed roles.*

Stereotypes as Cognitive Heuristics

Individuals are motivated to stereotype in order to cognitively understand (Fiske, 2000). Stereotypes are cognitive tools that can be efficiently employed to save mental effort through the application of categorical information onto targets based on highly salient membership cues, such as appearance or voice (Macrae, Milne, & Bodenhausen, 1994; Snyder, Tanke, & Berschied, 1977). Classical work on stereotyping suggests that stereotypes stem from a "need for coherence, simplicity, and predictability in the face of an inherently complex social environment” (p. 268, Tajfel, 1981). Stereotypes are distinct from prejudice that is often linked to them, in that they are generalized cognitive beliefs that are widely held throughout society about behaviors and attributes of individual members of social groups (Devine, 1989; Marx & Ko, 2019). Stereotype knowledge and application can be equally strong for individuals who have strong or weak prejudicial beliefs, and are a result of automatic, non-conscious activation of a well-learned set of associations developed through repeated exposure (Devine, 1989). Thus, the desire for female-voiced DVAs may not necessarily be rooted in prejudicial malice, but out of an ease of processing new stimuli.

Stereotypes as Nodes in a Cognitive Network

Stereotypes belong in a class of social schemas that aim to cognitively represent external reality (Hoffman & Hurst, 1990). The “Parallel-Constraint Satisfaction” theory (Kunda & Thagard, 1996) proposes that stereotypes, traits, and behaviors can be represented as interconnected nodes in a cognitive spreading activation network. The authors assume that nodes can both activate or deactivate each other, for example “red” and “truck” will facilitate recall of the closely cognitively-connected node of “firetruck”. Prior knowledge dictates the strength and direction of the connections between each node and its associates (Kunda & Thagard, 1996; Baltes, Bauer, & Frensch, 2007). Depending on what nodes are activated, individuals interpret stimuli differently. The same concepts, performed by different people or in different contexts, can have multiple interpretations (Sagar & Schofield, 1980). Thus, spreading node activation between cues (like voice gender) and pre-existent stereotypes help fill in missing information about a person or event and generate expectancies about what is likely to happen next (Hamilton, Sherman, & Ruvolo, 2010). These expectancies can serve as a guide to behavior during social interactions, so that a person can be ready to respond appropriately (Chen & Bargh, 1997).

There is much empirical evidence demonstrating the cognitive benefits that stereotypes can provide by applying pre-existing categorical information onto individual targets that show salient group cues. Most of the work that has quantified the cognitive benefits of stereotypes has shown their increased use in situations of high cognitive load. In the classic studies by Bodenhausen (1990; Bodenhausen & Lichtenstein, 1987), researchers demonstrate how decreased mental "energy" (due to cognitive constraints, like limited time) leads to increased reliance on efficient cognitive heuristics, like stereotypes. Similarly, Gilbert and Hixon (1991) show that cognitive load decreases the likelihood of stereotype activation, but increases the likelihood that an already activated stereotype will be applied. Even when individuals are highly motivated to not stereotype, high cognitive load can undermine this desire. Blair and Banaji (1996) experimentally manipulated intention not to stereotype and time constraints and found that those who created a counter-stereotype intention in advance of confronting new stimuli were able to decrease the automatic activation of gender stereotypes, but were unable to completely avoid them when under cognitive load.

Overall, the literature identifies stereotypes as a cognitive guide for what to expect in otherwise ambiguous situations, confronting novel stimuli by projecting characteristics and associations triggered by cues in the stimuli. Since stereotypes can help individuals figure out what to expect from a person or object based on their pre-existing mental model of concepts with the same associations, individuals are likely to have a greater sense of understanding of a novel device when it displays features that are easily associated with stereotypes.

In the case of DVAs, a novel user may confront a device and not know what to expect. Here, stereotypic cues can help the user figure it out. When the user hears the minimal cue of voice gender in a DVA, they activate all the concepts related to the voice gender, e.g., 'female'. Since one well-formed association is between women and support roles, like assistants (Rudman & Kilianski, 2000; Glick et al., 2005), the voice gender makes the concept of assistant more salient, which then facilitates the user's understanding by 'filling in the gaps' of their knowledge of the novel device with what they know about assistants in general. Thus, stereotype-congruent features could help users by providing them with a blueprint of what to expect from the novel device based on their prior knowledge of human equivalents.

Counter-stereotypes Facilitate Effortful Thinking

One counterargument to this spreading activation account of how gendered DVA voices aid users is found in the counter-stereotypes literature. Research has found that encountering a counter-stereotypical exemplar can encourage use of individuating information (Gocłowska et al., 2012; Hall & Crisp, 2005). In a priming study, gender counter-stereotypes enhanced cognitive flexibility, and social counter-stereotypes in general boosted creative performance (Goclowska, Crisp, & Labuschagne, 2013). Even for individuals who generally prefer to think less (i.e., low Need for Cognition), exposure to counter-stereotypes promotes cognitive reflection (Damer, Webb, & Crisp, 2019). Thus, encountering a counter-stereotype could actually induce more thinking (as opposed to imputing stereotypical features) about the device when presented with information about it for the first time, and therefore lead to greater understanding of its capabilities. According to this account, we should see that stereotype-incongruent exemplars lead to greater perceived understanding of the device's capabilities.

However, there are reasons to doubt that counter-stereotypes can successfully facilitate understanding of DVAs. One major challenge comes from the backlash caused from counter-stereotypes. The congruence literature has documented many instances in which counter-stereotypical features cause negative evaluations of the person or object of interest (Flannigan et al., 2013). Decreased liking of a DVA with counter-stereotypical features may cause potential users to be disinterested in adopting a new technology, even if they may understand its capabilities. Alternatively, the discomfort caused by stereotype-incongruent devices may lead to further disengagement, reduced exploration and use, and overall reduced understanding of the device. One mock trial study found that counter-stereotypical defendants caused jurors to become distracted by the defendant's counter-prototypicality, at the cost of paying less attention to the evidence presented (McKimmie et al., 2013). Although both accounts hold promise, I posit that the primacy of the activation of stereotypes would trump the more effortful processing inspired by counterstereotypes (Devine & Monteith, 1999; Bargh & Chartran, 1999). I therefore hypothesize that stereotype-congruent features in DVAs will help individuals actually understand the device better than stereotype-incongruent devices. Thus, my second hypothesis is:

H2: Individuals will demonstrate greater actual and perceived understanding of features for devices that are gender stereotype-congruent, with individuals feeling more certain about the capabilities of a female-voiced DVA in a female-typed job context, and male-voiced DVAs in male-typed roles, compared to male-voiced DVAs in female-typed roles or female-voiced DVAs in male-typed roles.

Overview of Studies

In Study 1, I show that individuals, on average, prefer a voice that is congruent with the stated job role's gender stereotype, across multiple job roles. In Study 2, I conceptually replicate this finding by presenting individuals with one job role, framed either in a stereotype-congruent or incongruent manner, and find that individuals again rate stereotype-congruent voices as more suitable. In Study 3, I extend these findings to show how these stereotypic preferences facilitate cognitive understanding of the device's capabilities.

Study 1

Although industry manufacturers suggest that there is an overall preference for women's voices in digital voiced agents (DVAs), Hypothesis 1 asserts that technology users will exhibit gender-role congruence by matching DVA voice gender to the gender-typing of the task. To test whether individuals actually prefer female voices in general or whether they prefer gender-role congruent voices for these personal devices, I asked participants in a pre-registered study to select which voice (either male or female) they prefer across six jobs that varied in gender-typing (3 male-typed, 3 female-typed).

Method

Participants

738 US residents were recruited from Amazon Mechanical Turk in return for market-rate compensation (M_{age} = 37.3, 50% women). As pre-registered in my exclusion criteria, I excluded respondents who had taken the survey twice (*N* = 20), did not give consent (*N* = 4), or failed one of the four basic attention checks. The first page of the survey included a sound test to make sure that the participants could hear the audio files present throughout the survey. Participants (*N* = 95) who could not correctly write down the number spoken on the first page were not

allowed to continue with the rest of the survey. The second check was immediately after hearing the first audio file introducing the digital voiced agent. Participants answered the question "To make sure that you were paying attention, what job does this AI agent do?" using a set of three multiple choice answers -- "Fact Checker", "Data Scientist", and a dynamic field that contained the correct answer of the actual job role that the participant was randomized into. 116 participants failed the second attention check. The last two attention checks were directly after the focal choice between the two voices, and asked about the content of the audio files. The two attention checks were preceded by an explanation "We also want to make sure that you listened to the audio files before. Please tell us the gender of the voice that you chose" (Answers: Male/Female; *N* = 32 failed) and the last question asking "What did the voices say?", which could be answered by selecting the one of three multiple choice answers that contained the exact text that was spoken in the sound file (*N* = 39 failed). Incomplete survey responses were not counted towards the total participant number, leaving 413 participants in the final dataset for analysis. The large (56%) number of participants that failed the attention checks reflects the poor quality of responses that affected many online crowdsourcing platforms due to the Covid-19 pandemic (responses were collected in the Summer of 2020; Arechar & Rand, 2020).

Pre-Tests

Job roles. conducted pre-tests of the perceived gender-typing and status of a set of jobs in order to generate stimuli for the main study. Using 32 job examples taken from research articles quantifying perceived occupational gendering (e.g., Heilman & Saruwatari, 1976; Ridgeway, 2011) as well as from the U.S. Census, I asked two separate samples of U.S. based participants, through Amazon's Mechanical Turk, to rate jobs on their perceptions of gender-typing (sample 1, *N* = 205) and status (sample 2, *N* = 220). Participants practiced with three example jobs where they were shown how to answer using the question format. After mastering the three examples, they were presented with 32 jobs serially in randomized order and answered how 'most people' would view that job in terms of gender-typing or status. They answered using a 5-point Likert scale ranging from 1 = Masculine to 5 = Feminine, or from 1 = Low status, to 5 = High status, respective to the survey.

I selected six jobs according to perceived gender-typing and status, such that there were three masculine and three feminine jobs that were matched for high, medium, and low status. Three jobs that varied in status but are similar in perceived femininity are Nurse (high status), Secretary (medium status), and Housekeeper (low status). For the masculine jobs, Janitor (low status) shared equivalent perceived status as Housekeeper, Security Guard (medium status) with Secretary, and Pilot (high status) with Nurse. Wilcoxon Signed-Ranks tests indicated that Nurse and Housekeeper were viewed significantly differently in terms of status (M_{nurse} = 4.12, SD = 0.72; M_{house} = 1.53, *SD* = 0.79; *Z* = -12.1, *p* < 0.001), as well as Pilot and Janitor (M_{pilot} = 4.48, *SD* = 0.67; $M_{janitor}$ = 1.42, *SD* = 0.82; *Z* = -12.2, *p* < 0.001). Similarly, they were significantly different in terms of gendered association (Housekeeper and Janitor, M_{house} = 4.54, *SD* = 0.76; $M_{janitor}$ = 1.80, *SD* = 0.89; *Z* = -11.8, *p* < 0.001; Secretary and Security Guard, $M_{secretary}$ = 4.44, *SD* = 0.82; M_{guard} = 1.33, *SD* = 0.82; *Z* = -12.1, *p* < 0.001; and Nurse and Pilot, M_{nurse} = 4.47, *SD* = 0.76; M_{pilot} = 1.81, *SD* = 0.83; *Z* = -11.7 , *p* < 0.001).

Voices. I synthesized a set of 15 voices using a Natural Language Processing Text-to-Speech generator that included a range of 'male' and 'female' voices with American accents. I asked 99 U.S. based participants through Amazon's Mechanical Turk to rate one randomly assigned voice on five traits deemed relevant in voice research (friendly, familiar, pleasant, assertive, smart) and a gender check of how masculine/feminine the voice sounds (Vinney & Vinney, 2017; Burgoon, 1978). I chose the two voices (one male and one female) that were rated as equivalent on all of the traits, as confirmed by a MANOVA (Pillai's Trace = 0.03, $F(1, 61) = 0.36$, $p = 0.87$), except for the gender check of perceived masculinity/femininity ($t(39) = -15.8$, $p < 0.001$), with the male voice being rated as significantly more masculine ($M = 1.6$, $SD = 1.1$) than the female voice ($M = 4.8$, $SD = 0.4$).

Procedure

Participants first read an introduction explaining that researchers at MIT were working on the next wave of commercial AI agents and that the purpose of the survey was to collect opinions on what personalities the participants liked best for these digital agents. The participants then were introduced to a specific AI agent that had a randomly assigned job. There were no images of the AI agent throughout the survey. The job was one of the six jobs selected from the pre-test, as described above. Participants then answered an attention check question on the job of the AI agent. I then told remaining participants that the AI agent was programmed with the basics of how fulfill its job role and that it updates its knowledge by interacting with real people in the workplace. I ensured that participants paid attention to the job role by asking them to write down three tasks, in open-ended text boxes, that a human occupying the job role would do. After this task, I told participants that the AI agent learns by talking to and listening to the people that it works with and presented the focal question of the survey: "We want your opinion on the best voice that suits it [the device]. Which voice do you prefer [for the device]?" Participants had two media players, one labelled Voice X (with the male voice) and one labelled Voice Y (with the female voice). After making their choice, I asked participants, on the next page, to explain their choice through an open-ended text box followed by the question “What influenced your choice? Why did you prefer the voice you selected for the job?”. Participants then answered more attention check questions (as described above). Next, I probed the participants’ beliefs on differential gendered abilities in the job they were randomly assigned, with a single slider that ranged from “Women are better” on one end and “Men are better” on the other end, and “No difference” in the middle to respond to the question “On average, how much do you think that women or men are better at [job role]?”. The scale was intentionally unnumbered as individuals do not have a quantified expectation of difference between women’s and men’s abilities, but rather a generalized subjective comparison of which gender would be better than the other, if at all. In order to compute difference scores, I translated the graphical slider scale to numerical equivalents, with -1 representing the polar end of “Women are better” and +1 representing “Men are better”. As with all of the surveys in this chapter, I collected participant gender, age, political affiliation, and employment status.

Results

As predicted, and shown in Figure 1 below, individuals chose the voice whose gender matched the stereotyped gender of the job presented (main effect of feminine job type on likelihood of choosing female voice: $B_{female\ job}$ =1.06, *SE* = 0.2, *p* < 0.001, 95% CI [0.66, 1.46]).

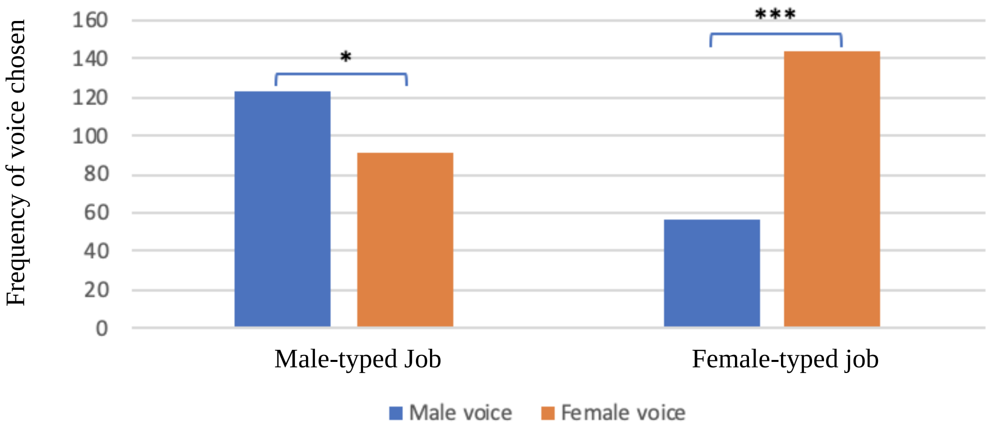


Figure 1. Frequency of voice chosen by gender-typing of job.

The choices broken down by job category reveal an overall pattern of gender stereotyping, as seen in Figure 2 below. On the whole, for female-typed jobs, participants preferred the female voice more than the male voice. The opposite trend held for male-typed jobs, where participants chose the male voice over the female voice on average. The only significant exception to this trend was for the job of Pilot (highest status, male-typed job), for which more participants preferred female voices than male voices. Review of the text responses where participants could justify their voice selection ("Why did you choose this voice?") showed that individuals were concerned about gender representation for the role of Pilot. Interestingly, gender representation for the other jobs was not a factor, suggesting an avenue for future research examining the differential concerns surrounding diversity for jobs of varying status and gender representation. The other partial exception was that the difference in voice preference for the job of Janitor was only marginally significant (although it is significant with a one-tailed test, given that it is a pre-registered hypothesis).

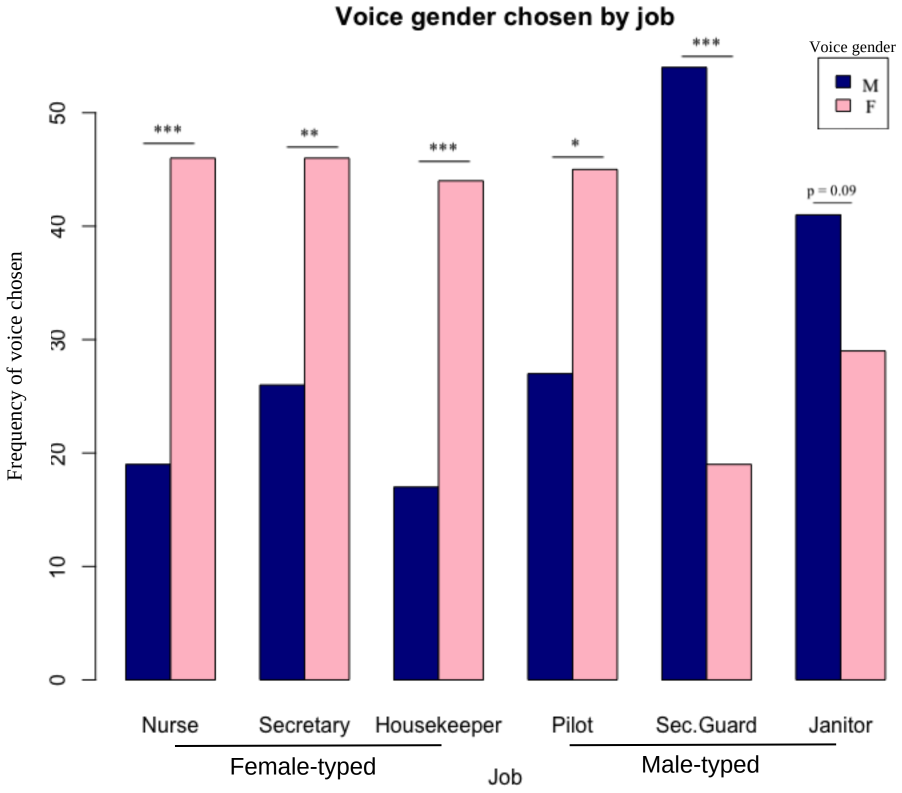


Figure 2. Frequency of voice chosen by gender-typing of specific job roles.

Discussion

Study 1 showed promising results that suggest that instead of a generalized preference for women's voices for DVAs, individuals show a preference for voices that are gender-role stereotype congruent. A potential limitation of this study is that the voices were matched on their pleasantness, familiarity, and other superficial characteristics. A male voice showing an equivalent amount of warmth and other positive vocal characteristics to a female voice may be an unrealistic comparison. However, if anything, this is a conservative test given the results supporting stereotype congruence. Even in this unnatural setting where a male voice chosen to be pleasant and warm could provide an additional boost to preferability, participants nevertheless chose the female voice for the female-typed job roles.

A second limitation of Study 1 is that the focal question of voice preference placed an emphasis on ‘the job’ and may therefore have created demand from the participants by revealing my interest in the role of stereotypes associated with the job. I address this limitation in Study 2 by improving the question wording, removing any explicit mention of the job role itself.

Given the promising results of this initial foray into establishing a gender-role stereotype congruence effect in driving the preferred vocal gender for DVAs, I designed Study 2 to conceptually replicate the relationship and improve the wording of the vocal preference question.

Study 2

The aim of Study 2 was to replicate the findings of Study 1 using different experimental stimuli and to assess whether individuals, when confronted with a specific voice-job pairing, would find voices in stereotype-incongruent pairings less suitable than stereotype-congruent ones. Instead of manipulating the job itself, as in Study 1, I held the job (of baking) constant and manipulated the gendered framing of the agent.

Method

Participants

1011 US residents were recruited from Amazon Mechanical Turk in return for market-rate compensation (M_{age} = 38.3, 55.3% women). I excluded any participants who did not consent (*N* = 1), had taken the survey before (as evidenced by repeated IP addresses, *N* = 2), failed the initial audio check before randomization into any condition (*N* = 12), failed one of the basic attention check questions (as elaborated in the methods section of Study 1, *N* = 27), or could not correctly identify Riley's voice gender (*N* = 43), leaving 938 responses. The improved rate of retention can be attributed to the use of data quality filter provided through CloudResearch for additional payment.

Procedure

After giving informed consent, all participants went through an audio manipulation check where they had to correctly type out a numerical code that they heard. Those who correctly wrote out the number then read a marketing blurb introducing them to a digital voiced agent designed to make baking at home more accessible to individuals. Participants were randomly assigned to have a feminine job or masculine job role frame. The masculine framing condition read:

> *There's nothing more scientific than baking, which requires careful measurement, a good understanding of chemistry and physics, and handling high temperatures. … Riley is built on the expertise and intelligence of many professional bakers.*

Whereas the feminine framing read:

> *There's nothing better than the comforting smells of fresh baked goods from the oven at home, to make you feel cozy and warm. … Riley is built on the tips and tricks of generations of home bakers.*

The masculine framing was designed based on prior research on gendered characteristics, where STEM topics (Young et al., 2013; Smeding, 2012), expertise (Gálvez, Tiffenberg, & Altszyler, 2019), and professionalism (Duehr & Bono, 2006; Ruiz Ben, 2007) are associated more with men than with women. The female framing was similarly based on research that links the home and homemaking (Eagly & Steffen, 1984; Eagly, Wood, Diekman, 2000) and warmth (Fiske et al., 1999; Decolette et al., 2013) with women. The framings each had one photo of bakers corresponding to the gendered framing (men baking in the masculine condition or a woman in the feminine condition).

After reading through the gendered framing, the participants were introduced to Riley, a gender-neutral name used in other personified AI research (Jago, Carroll, & Lin, 2021). Participants were randomly assigned to either a male or female voice for Riley. Participants encountered a photo of a generic device (as seen in Figure 3) and listened to Riley introduce itself. Riley said:

Hi, my name is Riley. I am an artificially intelligent baking ***assistant (instructor)****.*

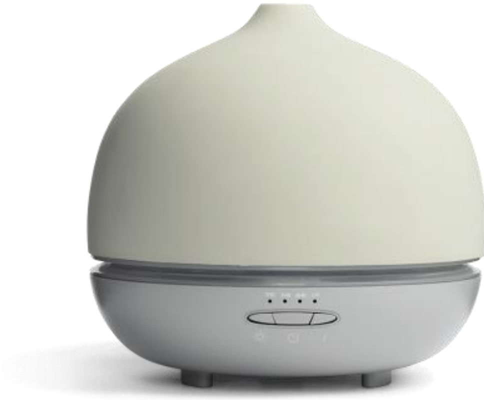


Figure 3. Stimuli from the experimental survey of Riley the fake AI baking agent.

Although there was no way of forcing the participants to listen to the sound clips, they were unable to advance in the survey until the time taken for the sound clips had completely elapsed, and there were attention check questions based on the content of the sound clips that prevented advancing through the survey, for example, participants had to successfully confirm Riley's topic area of expertise (baking). Afterwards, they listened to another audio clip where Riley described its capabilities. In the masculine condition, Riley said:

*I will **direct** you through the steps of baking, **supervising** your progress, and **tell** you what to do along the way*.

In the feminine condition, Riley said:

*I will **guide** you through the steps of baking, **supporting** your progress, and **answering all of your questions** along the way*.

On the next page, participants rated the suitability of Riley's voice to its intended purpose by answering "How well do you think Riley's voice suits its role?" on a 0-100 slider scale ranging from "Totally unsuitable" to "Perfectly suitable". I used suitability as equivalent to preference for the voice in this context (as measured by liking in Study 1) that read naturally within the question prompt.

The survey ended with a question probing the participants' beliefs on differential gendered abilities in baking in the same way as Study 1. On the whole, participants rated women as being slightly better at baking (*M* = -0.2, *SD* = 0.3). Finally, I asked participants to identify the gender of Riley's voice, where respondents could answer Female, Male, or "Couldn't tell" without penalty. Twenty-six participants responded that they "couldn't tell" the gender of the voice, and were dropped from the analyses.

Results

Manipulation Check. The framing manipulation was effective in creating a significant difference in perceptions of gendered associations in baking, *t*(936) = -3.2, *p* = 0.002. Participants who received the female framing manipulation rated women as being significantly better at baking (*M* = -0.21, *SD* = 0.3) compared to those who received the male frame (*M* = -0.14, *SD* = 0.3).

As predicted, and shown in Figure 4 below, individuals rated the voices as being more or less suitable dependent on gender-role congruence (Interaction effect of job role and voice gender on suitability rating: $\beta = 10.2$, $SE = 4.26$, $p = 0.017$). Unlike my prediction, and unlike the results in Study 1, the interaction effect was driven solely by the difference in ‘suitability’ ratings in the masculine “Instructor” role, for which the female voice was rated less suitable than the male voice.

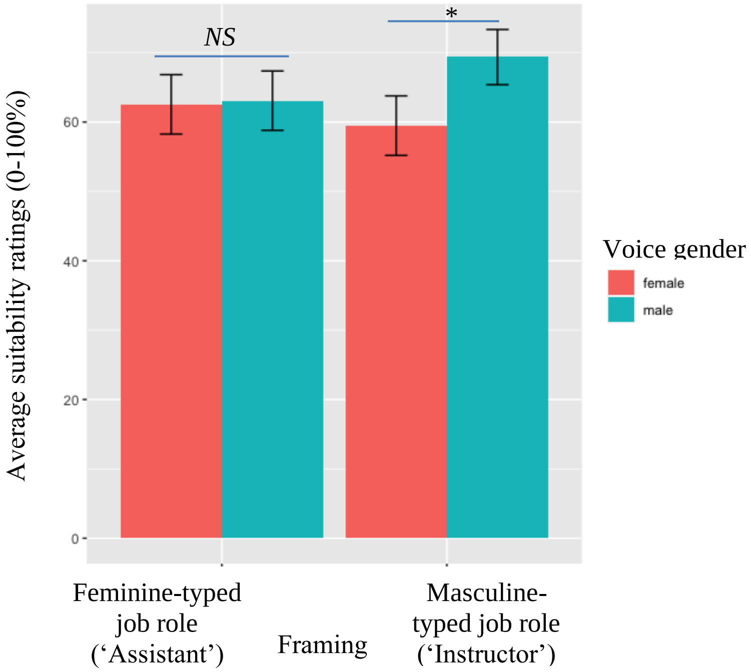


Figure 4. Average suitability ratings of voice by randomly assigned gender and job role. Error bars represent 95% CI.

Discussion

As hypothesized, there was a significant overall interaction effect of voice gender and job role congruence on evaluations of voice suitability, with congruent pairings rated as more suitable than incongruent ones, on the whole. Surprisingly, this was driven by the male gender-typed role, where the male voice was seen as much more suitable than the female one. Despite extensive research showing a backlash effect for men in female-typed roles (see Rudman, 2001, for a review), which was found in Study 1, there was no evidence of any penalization of the male-voiced DVA in the assistant role. Other gender research has shown that individuals change the meaning of a role depending on the person fulfilling it (Ramos et al., 2012). The Shifting Standards Model (Biernat, 1994) has also documented how the standards for evaluating women are different than the threshold for acceptability for men. In this study, female voices in the feminine role may have been held to a higher standard than the male voice in the atypical role. The absence of penalization of the male-voiced device in the role of assistant remains a puzzle that merits further investigation.

Study 2 was also limited in that the status elements of the job roles may have affected the evaluation of the voices, where ‘Instructor’ holds higher status associations than the role of ‘Assistant’. Gendered roles are difficult to match naturalistically on status due to persistent gender discrimination. Whereas there are myriad high-status male-typed jobs in the real world, there are very few professions with high-status, female-typed equivalents. Study 3 sought to ameliorate this issue by manipulating the gendered framing of the role context and capturing perceived status as a control variable. Further, Study 3 sought to test Hypothesis 2 that gender-stereotype congruent features serve a facilitatory role in helping the user understand an otherwise novel device.

Study 3

Given the facilitatory nature of schemas in processing new information, the literature on categorical schema activation would suggest that stereotype-congruent pairings ease understanding of a novel device’s capabilities (Rubin, Paolini, & Crisp, 2010; Macrae, Milne, &

Bodenhausen, 1994). However, work on exposure to examples of stereotype incongruence (Randsley de Moura et al., 2018; Goclowska, Crisp, & Labuschagne, 2013; Prati et al., 2015) has also demonstrated that counter-stereotypes inspire more effortful processing, increased use of individuating information, and a decreased reliance on heuristic judgment, suggesting that individuals may be more thoughtful when confronted with a novel device that possesses counter-stereotypic features.

Given the preference for gender-role congruent DVAs demonstrated in Studies 1 and 2, this study sought to test whether stereotype-congruent features help facilitate understanding of the device. In this 2 x 2 (Voice gender; Role gender framing) between-subjects design, I introduced participants to a novel digital voiced agent and measured how much the users understood the device's capabilities. Whereas Study 2 presented two distinct roles that varied in their gendered associations, Study 3 improved the design by holding constant the device's role and experimentally manipulating the gendered associations to be more male- or female-typed by presenting a neutral role within a gendered framing of the job role's broader domain. Study 3 also improved on the previous two studies by collecting information about participants' DVA usage: those who were more experienced with DVAs in general may already have had strong associations of DVAs with female characteristics, as the most commonly-used DVAs are female (CouponFollow, 2020). In addition to collecting participant information about their current DVA usage, I also chose a novel topic area for the DVA's purported role in order to minimize prior associations that participants could draw upon to understand the device.

The study consists of four conditions: two gender-stereotype congruent (male-voiced, male framing of role; female-voiced, female framing of role), and two incongruent (female-voiced, male framing of role; male-voiced, female framing of role).

As elaborated in the chapter introduction, I predict that participants will have a stronger mental model of stereotype-congruent exemplars to help them understand these novel devices. As a result, participants exposed to congruent (vs. incongruent) DVAs will be able to answer correctly a greater number of questions about the device's capabilities, be more certain about their understanding, and have a greater perceived understanding of the device.

Method

Participants

680 US residents were recruited from Amazon Mechanical Turk in return for market-rate compensation (M_{age} = 42.1, 51.6% women). As pre-registered in my exclusion criteria, I excluded 3 respondents who had taken the survey twice. One participant failed the attention check asking "To make sure that you were paying attention, please tell us what topic area Riley knows about" using a set of three multiple choice answers ("Baking", "Music", "Weather"). Fourteen participants failed the initial audio check, ten participants guessed the intention behind the study, 31 participants incorrectly identified the gender of the device's voice at the end of the survey, and one participant did not consent, leaving 620 responses for analysis.

Dependent Measures

Understanding. In order to quantify how much facilitation of knowledge occurred due to the presence of stereotype-congruent device features, I asked participants to answer five questions about the device's capabilities (e.g., "Riley can go through the steps of a recipe with you"), using three options: True, False, or Not sure. The five questions were based on both the written and voiced information already provided, thus every participant could have answered all the questions correctly. To measure actual understanding, I summed the total number of questions answered correctly (as True). To measure certainty in understanding, I summed the number of questions answered as other than "Not sure" (i.e., as definitively True or False). On average, participants answered 4.2 (*SD* = 0.9) questions correctly, and were sure of 4.4 (*SD* = 0.8) out of Riley's five features. As a qualitative measure of perceived understanding, I also asked participants "How well do you understand what Riley does?" on a slider ranging from "0 = I have no idea" to "100 = I understand fully". The average self-rated perceived understanding of Riley was 74% (*SD* = 20.3). As would be expected, actual and perceived understanding were positively correlated ($r = 0.86$, $p < 0.001$). Also as expected, the number of questions answered definitively and self-rated perceived understanding were significantly positively correlated ($r = 0.32$, $p < 0.001$).

Perceived Status. In order to control for possible differences in the perceptions of status caused by the gendered framings of baking, I asked participants to rate their perceptions of baking by answering the question "How high or low status do you perceived baking to be?" Participants answered on a 5-point Likert scale (Very low status, Low status, Neither low nor high status, High status, Very high status), where higher scores denote greater status. On average, participants across both gendered framings rated baking as a little above the middle of the scale ($M = 3.4$, $SD = 0.6$), and there was a significant, yet slight, difference in perceived status of baking between the feminine and masculine gendered framings ($t(611) = -2.4$, $p = 0.02$), with participants who watched the video of the female baker rating baking 0.1 points higher in status (on a 5-point scale) than those who viewed the video of the male baker. Accordingly, perceived status was controlled for in all models.

DVA Usage. Near the end of the survey, but prior to answering demographic questions, participants answered a Yes/No question asking "Do you use a voiced digital assistant (like Alexa, Google Assistant, Siri and others)? This could be through a smartphone or on a smart speaker or device."

Manipulation Check. Similar to the ratings found in Study 2, participants rated women as being slightly better at baking on average (*M* = -0.12, *SD* = 0.3, *t*(619) = -10, *p* < 0.001, *95% CI* [-0.14, -0.098]). Participants who received the female framing manipulation rated women as being significantly better at baking (*M* = -0.14, *SD* = 0.3) compared to those who received the male frame (*M* = -0.10, *SD* = 0.3; *t*(618) = -2.0, *p* = 0.05).

Procedure

After giving consent and passing the audio check test, participants were randomly assigned to either a masculine or feminine framing of baking. Both framing manipulations consisted of the first 50 seconds of a video advertisement featuring a famous French baker for online baking classes by Masterclass. The video featured either female baker Apollonia Poilâne or male baker Dominique Ansel. After the video, participants read the same gendered frame of baking as in Study 2, that corresponded to the video watched. Similar to Study 2, participants listened to Riley introduce itself through an audio file underneath a photo of a generic speaker-like device. Participants were randomly assigned to listen to a device with either a male or female voice. In all conditions, the content of what Riley said was held constant:

Hi, my name is Riley. I'm an artificially intelligent baking partner.

On the next page, after an attention check question, participants could listen to Riley give more information on its capabilities. In the Assistant condition, the text preceding the sound read "Riley is built on the tips and tricks of generations of home bakers", whereas in the Instructor condition, it read "Riley is built on the expertise and intelligence of many professional bakers". Participants then answered how suitable they thought the voice was for Riley, and afterwards answered the questions that measured their understanding of the device.

Next, participants gave their subjective evaluation of their understanding of Riley's capabilities. This was followed by the manipulation check question on the gendered framing presented at the beginning of the survey and, on the next page, a question about the perceived status of baking. Participants then answered an attention-check question about the voice gender of Riley. After this, I asked participants whether they used a voiced digital assistant (like Alexa, Google Assistant, Siri, Bixby) in a yes/no question that dynamically displayed follow-up questions on the characteristics of the voice if participants responded affirmatively. Finally, participants gave demographic information and answered an open-ended text box question asking for their best guess about the purpose of the survey. A research assistant reviewed these responses to flag any responses that suggested that the participant correctly identified the intention of the study, with the criteria being mention of both: a) the gender of the device and b) baking being a gendered domain. Ten participants (1.47%) correctly guessed the intention behind the survey and were eliminated from the study. I analyzed the data both including and excluding these participants; the focal effect remained significant and substantively unchanged.

Results

The main aim of Study 3 was to test whether devices with stereotype-congruent features would lead to greater understanding of the device's capabilities. In order to check that there were no idiosyncratic factors of the gendered framing affecting generalized understanding, I analyzed the average number of correct answers by those who received the masculine (vs. feminine) gendered framings. As expected, there was no difference in the average number of questions answered correctly for those who read the female (vs. male) framing manipulations ($t(623) = -1.1$, $p = 0.3$), suggesting that the gendered framings did not cause unintended differential facilitation of understanding the device.

When I subjected the data to a OLS regression as pre-registered with actual understanding as the dependent variable, I found that, although the means appeared in the hypothesized direction (see Figure 4), the predicted interaction effect between voice gender and gender framing of baking failed to reach significance ($p = 0.2$, as shown in Model 1 of Table 1). This pattern held for certainty in understanding ($p = 0.5$, in Model 1 of Table 2, see Figure 5) and the self-reported measure of perceived understanding ($p = 0.6$, in Model 1 of Table 3; Figure 4.3). However, since this sample consisted of both individuals who report using female-voiced digital assistants as well as novices, some of the participants may have had strong pre-existing expectations surrounding what a digital voiced assistant should sound like (i.e., female-voiced). I therefore subjected the data to a 2 x 2 x 2 ANOVA of voice gender and gendered role framing with the addition of an individual's prior usage of a DVA. There was a significant three-way interaction such that the effect of stereotype congruence on actual understanding differed for those who had used DVAs compared to those who had not ($B = 0.72$, $SE = 0.31$, $p = 0.02$, 95% CI: [0.11, 1.33] with demographics, see Models 3 & 4 in Table 1). Stereotype congruence was statistically significant for those without prior exposure to a DVA but not for those with prior exposure, as shown in Figure 4.3. This three-way interaction replicated for perceived understanding with marginal significance (*B* = 0.55, *SE* = 0.29, *p* = 0.06, 95% CI: [-0.01, 1.11], see Models 3 & 4 in Table 1), and did not reach significance for certainty (*B* = 2.83, *SE* = 7.1, *p* = 0.7, see Models 3 & 4 in Table 1).

To further investigate the effect of stereotype congruence on understanding, I separately analyzed the participants who had reported not being users of a DVA. Those exposed to stereotype-congruent conditions answered significantly more items about the device's capabilities correctly than those in incongruent conditions ($B_{Framing\ x\ voice}$ = 0.69, *SE* = 0.29, *p* = 0.02, 95% CI: [0.14, 1.24], as shown in Models 1 & 2, in Table 4), consistent with my hypotheses. As shown in Figure 4.3, analysis of the simple effects showed that in the masculine framing condition, participants who listened to the male-voiced device gave more correct answers (*M* = 4.3, *SD* = 0.9) than those exposed to the female-voiced device (*M* = 3.8, *SD* = 1.2; *t*(80) = -2.4, *p* = 0.02). In contrast, for those randomly assigned to the feminine condition, there was no significant difference in the number of questions answered correctly between participants who had heard the female-voiced device compared to those who heard the male-voiced device (*p* = 0.5).

There was no main effect of participant gender (all *p*'s > 0.1) for the effect of stereotype congruence on actual, perceived, or self-reported understanding, as seen in Tables 1-4, for all subjects. The inclusion of a full range of demographic control variables (age, race, political affiliation, and employment status) did not substantively change the effect of stereotype congruence for all three forms of understanding (see Tables 1-4).

Discussion

Study 3 demonstrated that the congruence between stereotypical associations of a job role and the voice of the device operating within that role matter for actual understanding, but only if the users lack pre-existing expectations of a DVA. When users without prior experience were presented with a device that had a voice gender that matched the gender-typing of the role that it possessed (i.e., a device with stereotype-congruent features), users were able to answer a greater number of questions correctly as compared to when the device had features incongruent with the gender-framing of the role.

This effect was largely driven by congruence in the male-frame condition, where participants demonstrated greater understanding of the device's capabilities when the voice gender was congruent rather than incongruent with the gender typing of the role. Whereas male features in male-typed roles are lauded, female features that are incongruent with the male associations of the job role are not only penalized (as in Studies 1 & 2), but are less well understood.

Surprisingly, participants were not explicitly aware of the facilitatory effects of congruence, as noted by the lack of effect on their certainty of understanding of the device's capabilities. These findings highlight the implicit nature of stereotype congruence, where the facilitatory effects of congruent devices are below the threshold of awareness for individuals. Thus, researchers and policy makers should seek a better understanding of the role of stereotype congruence in influencing preference for stereotypical features in smart consumer devices and ascertain whether the cost of exacerbating stereotypical tropes outweighs the benefits to comprehensibility and the adoption of new technologies.

A limitation of this study is that it did not experimentally manipulate prior exposure to DVAs. Although I chose a novel topic area, where no consumer DVA exists currently, it is likely that the association between female voice and DVA is already entrenched for current users and therefore would supersede the more subtle benefits of stereotype congruence to facilitate understanding. It is likely that the effect of stereotype congruence was pronounced in non-users as they are less familiar with DVAs in general and are more likely to rely on their pre-existing stereotypes to understand the device. This is in line with previous research that shows that individuals are more likely to apply stereotypes when in unfamiliar situations (Bowles et al., 2005), and also when there is less individuating information available (Ginosar & Trope, 1980; Nisbett & Borgida, 1975).

A limitation of the methodology that may have caused the lack of effect for the feminine framing could be due to the masculine language spoken by Riley when describing its capabilities. Since the purported tasks were directive and masculine – traits incongruent with the female voice – participants may have penalized female-voiced Riley in the feminine frame because of the incongruence between her voice and her role. Similarly, male-voiced Riley may have benefited from the congruence between its voice and the masculine job tasks it described. This effect may have cancelled out the greater congruence effect of the initial gendered framing, and thus led to a null comparison between the male and female voice in the feminine job frame. A future study that controls the language that Riley uses to describe its features is recommended.

Although the generalizability of the current results must be confirmed in future research using different domain areas and a variety of voices, the present study has identified support for stereotype-congruent pairings being advantageous for designers of novel personified technologies and highlights a potential rationale for the prevalence of gender-stereotypic female-voiced DVAs.

General Discussion

In three experimental studies, I show that individuals, on average, prefer a DVA with a voice that is congruent with its stated job role's gender stereotype, across multiple job roles. This effect is largely driven by congruence in the male-frame condition; whereas male voices in male-typed roles are lauded, female voices that are incongruent with the male associations of the job role are not only penalized (as in Studies 1 & 2), but are less well understood (as found in Study 3). Stereotypic preferences facilitate understanding of the device's job role for novel devices. These results suggest that there are cognitive benefits to stereotype-congruent features embedded in DVAs. Thus, the preference for stereotypical devices may stem not only from a desire to maintain a gender hierarchy, but possibly due to a desire simply to understand a novel technology more easily. There are three specific elements of my findings that I will contextualize within the broader research literature: the exacerbation of the stereotype congruence effect in male-typed roles in Studies 2 & 3; the benefits to understanding being limited to novice users; and the lack of awareness of the benefits to understanding.

Stereotype Congruence is Exacerbated in Male-Typed Roles

In both Study 2 and Study 3, evaluations of the male-voiced DVA were significantly affected by stereotype congruence of the device's voice gender to the gender-typing of its role. However, there was no difference caused by stereotype congruence for the female-voiced DVA. This pattern is consistent with prior research that shows that men are penalized for deviations from gender stereotypes, sometimes to greater degrees than women, due to the threat posed by these deviations to status hierarchies (Rudman et al., 2012). Whereas women generally occupy low-status roles and their striving for greater status can be more readily understood, observers are harsher against men who appear to renounce their high status by behaving in gender atypical ways. For example, men – but not women – who asked for help were penalized with lowered ratings of competence, as they failed to live up to the stereotype of a male leader who is independent and self-reliant (Rosette, Mueller, & Lebel, 2015).

This pattern is also consistent with previous research that shows that women in male-typed industries are penalized more than men in female-typed domains. Prior work contrasts the "Glass ceiling" that prevents career success for women in masculine professions due to a mismatch between prescriptive female gender roles and the masculine characteristics demanded from job roles, to the "Glass elevator", where men are afforded benefits and privileges in female-dominated professions (Williams, 1992; Casini, 2016). This pattern of women in male-typed domains being penalized is seen in the results of this chapter where female-voiced DVAs are penalized with lowered ratings of suitability (in Study 2) and reduced comprehension (in Study 3) of the device's features when cast in male-typed roles, with female-voiced DVAs being understood less well by novice users when its domain area of expertise was cast as male-typed.

Interestingly, stereotype-congruence effects were found for both male- and female-typed roles found in Study 1, where participants directly compared a male and female voice against each other. One explanation for this difference could be the Shifting Standards Model (Biernat, 1994; Biernat & Fuegen, 2001). This theory states that individuals evaluate targets (e.g., female DVA voices) in comparison to their within-category peers (e.g., other women), as opposed to the broader pool of targets with dissimilar social characteristics (e.g., other women and men), and thus use differing standards for acceptability. For example, participants in Studies 2 and 3 may have been evaluating the male-voiced DVA in comparison to other men in that role, instead of as compared to both men and women in that role. In contrast, participants in Study 1 were forced to directly compare a male and female voice against each other and thus their comparisons may have more distinctly drawn out the effect of stereotype congruence that may have been obscured in Studies 2 and 3.

As a final point, one other interesting finding was that individuals chose the counter-stereotypic female voice for pilots – a high-status, male-typed role. This was the only job role where participants actively pointed to gender imbalance in the industry in their justifications for the DVA. These results may be evidence for how conscious awareness of stereotyping can be reversed when motivated by desire for more equitable representations. Interestingly, participants were not concerned about gender representation for the low-status roles, nor for the high-status, female-typed jobs. Future research should investigate whether individuals show a differential interest in gender equity for high-status roles only.

The Benefits to Understanding are Limited to Novice Users

Study 3's results suggest that the benefit of stereotype congruence to cognitive understanding is limited only to novice users that are unfamiliar with DVAs, and not to all users, as hypothesized. Stereotype congruence shows a modest facilitatory effect on understanding of novel products for individuals who lack a pre-existent mental model for them. The majority of the participants in Study 3 were users of digital voiced assistants, demonstrating their widespread adoption. Although I chose a novel job role (i.e., no DVA specific to baking is commercially available), it is likely that the association between female voice and DVAs in general is already entrenched for current users and therefore would supersede the more subtle benefits of stereotype congruence to facilitate understanding. The effect of stereotype congruence was pronounced in non-users as they are less familiar with DVAs in general and are more likely to rely on their pre-existing stereotypes to understand the device.

This is in line with previous research that shows that individuals are more likely to apply stereotypes when in ambiguous or unfamiliar situations (Bowles et al., 2005; Kunda & Sherman-Williams, 1993; Miles & LaSalle, 2008), and also when there is less individuating information available (Ginosar & Trope, 1980; Nisbett & Borgida, 1975). In classic work by Tajfel (1969), individuals who were less tolerant of ambiguity stereotyped more than those more comfortable with ambiguity. Thus, the benefits to cognitive understanding from stereotype-congruent pairings are likely to be limited to those who were unfamiliar with the novel device. The findings of this effect on non-users also gives us insight into how the first-ever audience of DVAs – when the concept of DVA was non-existent in the general public – used stereotypic features to augment their understanding of an otherwise-unknown device.

Individuals Lack Awareness of the Benefits to Understanding

Surprisingly, participants were not explicitly aware of the facilitatory effects of congruence. Stereotype congruence had no significant effect on perceived understanding of the device's capabilities, or certainty of that understanding. These findings highlight the implicit nature of stereotype congruence; the facilitatory effects of congruent devices are below an individual's threshold of awareness. A robust literature has demonstrated that individuals are unaware of how implicit associations and stereotypes shape their behaviors (see Dovidio & Gaertner, 2000, for a review). Much less work has experimentally demonstrated the awareness of the cognitive benefits of stereotypes, thus this chapter adds to the literature on how individuals can lack awareness of the effect of implicit associations on understanding new concepts (Greenwald & Banaji, 1995).

Theoretical Implications

These findings add nuance to the theoretical discussion on stereotype use by outlining the facilitatory benefits afforded by the use of stereotype-congruent features. Whereas status (i.e., Rudman, Moss-Racusin, Phelan, & Nauts, 2012) and system justification (i.e., Pratto & Espinoza, 2001) accounts of gender stereotyping have identified the maintenance of hierarchies in the perpetuation of stereotypes, I extend the research that focuses on stereotypes as cognitive shortcuts (Jussim et al., 1995). Individuals may prefer stereotype-congruent features not only due to prejudice or a desire to maintain social hierarchies, but simply because they are easier to understand. This is important theoretically: we must move the conversation away from focusing solely on individuals with prejudicial intent to highlighting how everyday individuals, including those who are highly motivated to avoid prejudicial behavior, can fall into preferring stereotype-congruent products (and people). It also raises a potential barrier to removing stereotype-congruent features in everyday devices: the cognitive efficiencies of stereotypical features, not prejudice, may be what encourages stereotype-congruent DVA design choices

Practical Implications

Despite the enormity of the voiced digital agent industry and user base that would be affected, calls for legislation to outlaw the default gendering of digital voiced assistants (Adams & Loideáin, 2019) and international policy guidelines are emerging without any empirical basis (Jobin, Ienca, & Vayena, 2019; West, Kraut, & Chen, for UNESCO, 2019). This work seeks to fill the gap in the literature by providing empirical evidence of the use of gender stereotyping in the features assigned to DVAs, and to theoretically examine the reasons for why individuals would prefer devices with features that are congruent with gender stereotypes.

I challenge the existing justifications for the default use of women's voices for DVAs as not merely a benign preference for women's voices, but as a persistent pattern of gender stereotyping. This identification of a systematic preference according to stereotypes is a consequential discovery as these hyper-scalable technologies have the potential of exponentially amplifying gender stereotypes. If women's voices were generally preferred for all devices equivalently, then the risk of exacerbating gender stereotypes would be much lower than if there was preferential selection of women's voices for roles expected to be for women only. This is an interesting societal implication of the work that needs further research to quantify the costs of gender stereotyping caused by novel devices.

Future Directions

No work to date has examined the role of gender stereotypes in guiding preference for digital voiced assistants, and thus this work demonstrates a novel application of gender stereotype congruence theory to interaction with personified algorithmic technology. However, the literature still lacks a comprehensive understanding of the potentially-detrimental effects of interacting with a gender-stereotype amplifying technology. From a societal standpoint, the ubiquity of voiced digital assistants, coupled with the exponential increase in artificial intelligent agents in everyday life, necessitates further research quantifying the effect of repeated exposure to gender-stereotypic representations on our stereotypes of humans. One avenue for research that should be given priority is naturalistic field experiments with actual devices of varying voice genders in the quantification of their effect on gender stereotyping of others. A large-scale, longitudinal field study of novice users of DVAs could compare the stereotypic associations before and after long-term interaction with stereotype-congruent or incongruent devices. Given the modest effects found in this chapter, it is likely that a repeated association that is sustained over a long period of time – akin to actual usage – would generate stronger and more realistic effects of gender stereotype congruence. Thus, in light of these results, researchers and policy makers should seek a better understanding of the role of stereotype congruence in influencing preference for stereotypical features in smart consumer devices and ascertain whether the cost of exacerbating stereotypical tropes outweighs the benefits to comprehensibility and the adoption of new technologies.

A Final Word: The Social Impact of Digital Technologies

In sum, individuals show preference for gender stereotypic features for DVA and this preference may be driven in part by the cognitive efficiencies afforded by stereotypes. Given the ubiquity of digital assistants, we need to examine how repeated exposure to their gender-stereotype congruence can impact our stereotypes of humans. Investigating the features of digital voice assistants is enlightening not only for those interested in technological adoption, but also for those invested in workplace diversity, as the future of work includes issues of representation of social identities of both human and digital actors. If scholars and practitioners alike are concerned about the ramifications of hiring one person on organizational diversity, we need to be equally, if not more, concerned about the representation put forward by this hyper-scalable technology. This work seeks to help propel current conversations around organizational diversity to consider not just the human but also the digital actors that help us do our work and the psychological impact they exert on our social worlds.

References

Adroit Market Research. (2020, February 24). *Intelligent Virtual Assistant (IVA) Market to grow at 33% CAGR during forecast period (2020-2025) - Insights on Growth Drivers, Size and Share Analysis, Key Trends, Leading Players, and Business Opportunities* [Report]. https://www.globenewswire.com/news-release/2020/02/24/1988963/0/en/Intelligent-Virtual-Assistant-IVA-Market-to-grow-at-33-CAGR-during-forecast-period-2020-2025-Insights-on-Growth-Drivers-Size-and-Share-Analysis-Key-Trends-Leading-Players-and-Busin.html

Arechar, A. A., & Rand, D. G. (2020). Turking in the time of COVID. Preprint available on PsyArxiv. https://doi.org/10.31234/osf.io/vktqu

Armstrong, C. L., & McAdams, M. J. (2009). Blogs of information: How gender cues and individual motivations influence perceptions of credibility. *Journal of Computer-Mediated Communication*, *14*, 435–456.

Baltes, B. B., Bauer, C. B., & Frensch, P. A. (2007). Does a structured free recall intervention reduce the effect of stereotypes on performance ratings and by what cognitive mechanism?. Journal of Applied Psychology, 92(1), 151.

Bargh, J. A., & Chartrand, T. L. (1999). The unbearable automaticity of being. *American psychologist*, *54*(7), 462.

Borkowska, B., & Pawlowski, B. (2011). Female voice frequency in the context of dominance and attractiveness perception. *Animal Behaviour*, *82*(1), 55-59.

Brown, B. L., Strong, W. J., & Rencher, A. C. (1973). Perceptions of personality from speech: Effects of manipulations of acoustical parameters. *The Journal of the Acoustical Society of America*, *54*(1), 29-35.

Bonnington, C. (2018, May 20). *Why It Matters That Alexa and Google Assistant Finally Have Male Voices*. Slate Magazine. https://slate.com/technology/2018/05/alexa-google-finally-are-getting-male-virtual-assistants-its-about-time.html.

Bowles, H. R., Babcock, L., & McGinn, K. L. (2005). Constraints and triggers: situational mechanics of gender in negotiation. Journal of personality and social psychology, 89(6), 951.

Bumby, K., & Dautenhahn, K. (1999, August). Investigating children's attitudes towards robots: A case study. In Proc. CT99, *The Third International Cognitive Technology Conference* (pp. 391-410).

Burgess, D., & Borgida, E. (1999). Who women are, who women should be: Descriptive and prescriptive gender stereotyping in sex discrimination.Psychology, Public Policy, and Law, 5(3), 665.

Burgoon, J. K. (1978). Attributes of the newscaster's voice as predictors of his credibility. *Journalism Quarterly*, *55*(2), 276-300.

Casini, A. 2016. Glass Ceiling and Glass Elevator. The Wiley Blackwell Encyclopedia of Gender and Sexuality Studies, First Edition. Edited by Nancy A. Naples. John Wiley & Sons, Ltd. Published 2016 by John Wiley & Sons, Ltd. DOI: 10.1002/9781118663219.wbegss262

Cialdini, R. B., & Trost, M. R. (1998). Social influence: Social norms, conformity and compliance. Chapter 21. In The Handbook of Social Psychology, ed. D. T. Gilbert, S. T. Fiske, G. Lindzey, 2:151–92. Boston: McGraw-Hill. 4th ed.

CouponFollow, 2020. https://couponfollow.com/research/voice-assistants-online-shopping

DeCasper, A. J., & Fifer, W. P. (1980). Of human bonding: Newborns prefer their mothers' voices. *Science, 208*(4448), 1174-1176.

DeLacollette, N., Dumont, M., Sarlet, M., & Dardenne, B. (2013). Benevolent sexism, men’s advantages and the prescription of warmth to women. *Sex Roles*, *68*(5-6), 296-310.

Devine, P. G., & Monteith, M. J. (1999). Automaticity and control in stereotyping.

Dovidio, J. F., & Gaertner, S. L. (2000). Aversive racism and selection decisions: 1989 and 1999. *Psychological science*, *11*(4), 315-319.

Duehr, E. E., & Bono, J. E. (2006). Men, women, and managers: are stereotypes finally changing?. *Personnel psychology*, *59*(4), 815-846.

Eagly, A. H., & Karau, S. J. (2002). Role congruity theory of prejudice toward female leaders. *Psychological review*, *109*(3), 573.

Eagly, A. H., & Johnson, B. T. (1990). Gender and leadership style: A meta-analysis. *Psychological bulletin, 108*(2), 233.

Eagly, A. H., Wood, W., & Diekman, A. B. (2000). Social role theory of sex differences and similarities: A current appraisal. *The developmental social psychology of gender*, *12*, 174.

Ellemers, N. (2018). Gender stereotypes. *Annual review of psychology*, *69*, 275-298.

Feinberg, D. R., Jones, B. C., Little, A. C., Burt, D. M., & Perrett, D. I. (2005). Manipulations of fundamental and formant frequencies influence the attractiveness of human male voices. *Animal* Behaviour, 69(3), 561-568.

Flannigan, N., Miles, L. K., Quadflieg, S., & Macrae, C. N. (2013). Seeing the unexpected: Counterstereotypes are implicitly bad. *Social cognition*, *31*(6), 712-720.

Fiske, S. T., Cuddy, A. J., Glick, P., & Xu, J. (2002). A model of (often mixed) stereotype content: competence and warmth respectively follow from perceived status and competition. *Journal of personality and social psychology, 82*(6), 878.

Fiske, S. T., Xu, J., Cuddy, A. C., & Glick, P. (1999). (Dis) respecting versus (dis) liking: Status and interdependence predict ambivalent stereotypes of competence and warmth. *Journal of social issues*, *55*(3), 473-489.

Fiske, S. T., & Glick, P. (1995). Ambivalence and stereotypes cause sexual harassment: A theory with implications for organizational change. *Journal of Social Issues*, *51*(1), 97-115.

Förster, J., Higgins, E. T., & Strack, F. (2000). When stereotype disconfirmation is a personal threat: How prejudice and prevention focus moderate incongruency effects. *Social Cognition*, *18*(2), 178-197.

Gálvez, R. H., Tiffenberg, V., & Altszyler, E. (2019). Half a century of stereotyping associations between gender and intellectual ability in films. *Sex Roles*, *81*(9), 643-654.

Glick, P., Larsen, S., Johnson, C., & Branstiter, H. (2005). Evaluations of sexy women in low- and high-status jobs. *Psychology of women quarterly*, *29*(4), 389-395.

Grand View Research. (2020, April). Intelligent Virtual Assistant Market Size, Share & Trends Analysis Report By Product (Chatbot, Smart Speakers), By Technology, By Application (BFSI, Healthcare, Education), By Region, And Segment Forecasts, 2020 - 2027[Report]. https://www.grandviewresearch.com/industry-analysis/intelligent-virtual-assistant-industry

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: attitudes, self-esteem, and stereotypes. *Psychological review*, *102*(1), 4.

Gregory, S. (1994). Sounds of power and deference: Acoustic analysis of macro social constraints on microinteraction. *Sociological Perspective*, 37, 497–526.

Gregory, S., Webster, S., & Huang, G. (1993). Voice pitch and amplitude convergence as a metric of quality indyadic interviews. *Language and Communication*, 13, 195–217.

Griggs, 2011. https://www.cnn.com/2011/10/21/tech/innovation/female-computer-voices

Hamilton, D., Sherman, S. & Ruvolo, C. (2010). Stereotype-Based Expectancies: Effects on Information Processing and Social Behavior. *Journal of Social Issues*. 46. 35 - 60. 10.1111/j.1540-4560.1990.tb01922.x.

Heilman, M. E., & Okimoto, T. G. (2007). Why are women penalized for success at male tasks?: the implied communality deficit. *Journal of Applied Psychology*, *92*(1), 81.

Klofstad, C. A., Anderson, R. C., & Nowicki, S. (2015). Perceptions of competence, strength, and age influence voters to select leaders with lower-pitched voices. *PloS one*, *10*(8), e0133779.

Ko, S. J., Judd, C. M., & Blair, I. V. (2006). What the voice reveals: Within-and between-category stereotyping on the basis of voice. *Personality and Social Psychology Bulletin*, *32*(6), 806-819.

Ko, S. J., Judd, C. M., & Stapel, D. A. (2009). Stereotyping based on voice in the presence of individuating information: Vocal femininity affects perceived competence but not warmth. *Personality and Social Psychology Bulletin*, *35*(2), 198-211.

Kunda, Z., & Sherman-Williams, B. (1993). Stereotypes and the construal of individuating information. *Personality and Social Psychology Bulletin*, *19*, 90–99.

Kuwabara, K., & Thébaud, S. (2017). When Beauty Doesn't Pay: Gender and Beauty Biases in a Peer-to-Peer Loan Market. *Social Forces*, *95*(4), 1371-1398.

Lee & Kisilevsky. https://onlinelibrary.wiley.com/doi/abs/10.1002/dev.21084

Marx, D., & Ko, S. (2019). Stereotypes and prejudice, in the Oxford Encyclopedia of Psychology. Retrieved April 20, 2021, from https://doi.org/10.1093/acrefore/9780190236557.013.307

Mayew, W. J., Parsons, C. A., & Venkatachalam, M. (2013). Voice pitch and the labor market success of male chief executive officers. Evolution and Human Behavior, 34(4), 243-248.

McKimmie, B. M., Masters, J. M., Masser, B. M., Schuller, R. A., & Terry, D. J. (2013). Stereotypical and counterstereotypical defendants: Who is he and what was the case against her? Psychology, Public Policy, and Law, 19(3), 343–354. https://doi.org/10.1037/a0030505

Miles, E. W., & LaSalle, M. M. (2008). Asymmetrical contextual ambiguity, negotiation self-efficacy, and negotiation performance. *International Journal of Conflict Management*.

Mitchell, W. J., Ho, C. C., Patel, H., & MacDorman, K. F. (2011). Does social desirability bias favor humans? Explicit–implicit evaluations of synthesized speech support a new HCI model of impression management. *Computers in Human Behavior*, *27*(1), 402-412.

Monroe, A. H. (2011). *Stereotyping based on within category cues: the effect of afrocentric features and vocal femininity on judgments of competence and warmth* (Doctoral dissertation, Sciences).

Moss-Racusin, C. A., Phelan, J. E., & Rudman, L. A. (2010). When men break the gender rules: Status incongruity and backlash against modest men. *Psychology of Men & Masculinity*, *11*(2), 140.

Nass, C., Moon, Y., & Green, N. (1997). Are machines gender neutral? Gender-stereotypic responses to computers with voices. *Journal of applied social psychology, 27*(10), 864-876.

Nass, C., & Yen, C. (2010). *The man who lied to his laptop: What we can learn about ourselves from our machines*. Penguin.

Neal, T. M. S., Guadagno, R. E., Eno, C. A., & Brodsky, S. L. (2012). Warmth and competence on the witness stand: Implications for the credibility of male and female expert witnesses. *Journal of the American Academy of Psychiatry and the Law, 40*, 488–497.

Pew Research Center. (2017, December 12). *Nearly half of Americans use digital voice assistants, mostly on their smartphones* [Report]. http://pewrsr.ch/2kquZ8H

Phelan, J. E., Moss-Racusin, C. A., & Rudman, L. A. (2008). Competent yet out in the cold: Shifting criteria for hiring reflect backlash toward agentic women. Psychology of Women Quarterly, 32(4), 406-413.

Phelan, J. E., & Rudman, L. A. (2010). Reactions to ethnic deviance: The role of backlash in racial stereotype maintenance. *Journal of personality and social psychology, 99*(2), 265.

Puts, D. A., Gaulin, S. J., & Verdolini, K. (2006). Dominance and the evolution of sexual dimorphism in human voice pitch. *Evolution and human behavior, 27*(4), 283-296.

Randsley de Moura, G., Leicht, C., Leite, A. C., Crisp, R. J., & Gocłowska, M. A. (2018). Leadership diversity: Effects of counterstereotypical thinking on the support for women leaders under uncertainty. *Journal of Social Issues, 74*(1), 165-183.

Rubin, M., Paolini, S., & Crisp, R. (2010). A processing fluency explanation of bias against migrants. Journal of Experimental Social Psychology, 46, 21-28.

Rudman, L. A. (1998). Self-promotion as a risk factor for women: the costs and benefits of counterstereotypical impression management. *Journal of personality and social psychology*, *74*(3), 629.

Rudman, L. A., & Fairchild, K. (2004). Reactions to counterstereotypic behavior: the role of backlash in cultural stereotype maintenance. *Journal of personality and social psychology*, *87*(2), 157.

Rudman, L. A., & Kilianski, S. E. (2000). Implicit and explicit attitudes toward female authority. *Personality and social psychology bulletin*, *26*(11), 1315-1328.

Rudman, L. A., & Glick, P. (2001). Prescriptive gender stereotypes and backlash toward agentic women. *Journal of social issues*, *57*(4), 743-762.

Rudman, L. A., Moss-Racusin, C. A., Phelan, J. E., & Nauts, S. (2012). Status incongruity and backlash effects: Defending the gender hierarchy motivates prejudice against female leaders. *Journal of Experimental Social Psychology*, *48*(1), 165-179.

Ruiz Ben, E. (2007). Defining expertise in software development while doing gender. *Gender, Work & Organization*, *14*(4), 312-332.

Sagar, H. A., & Schofield, J. W. (1980). Racial and behavioral cues in black and white children's perceptions of ambiguously aggressive acts. Journal of personality and social psychology, 39(4), 590.

Smeding, A. (2012). Women in science, technology, engineering, and mathematics (STEM): An investigation of their implicit gender stereotypes and stereotypes' connectedness to math performance. *Sex roles*, *67*(11), 617-629.

Sokhi, D. S., Hunter, M. D., Wilkinson, I. D., & Woodruff, P. W. (2005). Male and female voices activate distinct regions in the male brain. *Neuroimage*, *27*(3), 572-578.

Stern, C., West, T. V., & Rule, N. O. (2015). Conservatives negatively evaluate counterstereotypical people to maintain a sense of certainty. *Proceedings of the National Academy of Sciences*, *112*(50), 15337-15342.

Stern, J. (2017, February 21). Alexa, Siri, Cortana: The problem with All-Female digital assistants. Retrieved April 19, 2021, from https://www.wsj.com/articles/alexa-siri-cortana-the-problem-with-all-female-digital-assistants-1487709068

Tajfel, H. (1969). Cognitive Aspects of Prejudice. *Journal of social issues*, *25*(4), 79-97.

Terrell, J., Kofink, A., Middleton, J., Rainear, C., Murphy-Hill, E., Parnin, C., & Stallings, J. (2017). Gender differences and bias in open source: Pull request acceptance of women versus men. *PeerJ Computer Science, 3*, e111.

Vailshery, L. S. (2021a, January 22). Smart home devices supported by voice assistants in 2019, by product [Report]. https://www.statista.com/statistics/933551/worldwide-voice-assistant-supported-smart-home-devices/

Vailshery, L. S. (2021b, January 22). Number of digital voice assistants in use worldwide from 2019 to 2024 (in billions) [Report]. https://www.statista.com/statistics/973815/worldwide-digital-voice-assistant-in-use/

Vinney, C., & Vinney, L. A. (2017). That sounds familiar: The relationship between listeners' recognition of celebrity voices, perceptions of vocal pleasantness, and engagement with media. *Journal of Radio & Audio Media*, *24*(2), 320-338.

Wigboldus, D. H., Dijksterhuis, A., & Van Knippenberg, A. (2003). When stereotypes get in the way: Stereotypes obstruct stereotype-inconsistent trait inferences. Journal of Personality and Social Psychology, 84(3), 470.

Williams, C. (1992). The Glass Escalator: Hidden Advantages for Men in the "Female" Professions. *Social Problems, 39*(3), 253-267. doi:10.2307/3096961

Young, D. M., Rudman, L. A., Buettner, H. M., & McLean, M. C. (2013). The influence of female role models on women's implicit science cognitions. *Psychology of Women Quarterly*, *37*(3), 283-292.

Figures

Figure 1. Frequency of Voice Chosen by Gender-Typing of Job

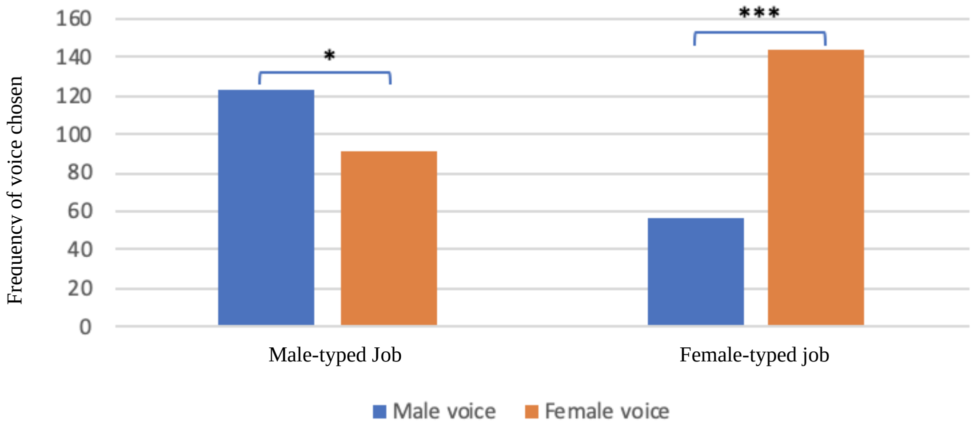


Figure 2. Frequency of Voice Chosen by Gender-Typing of Specific Job Roles

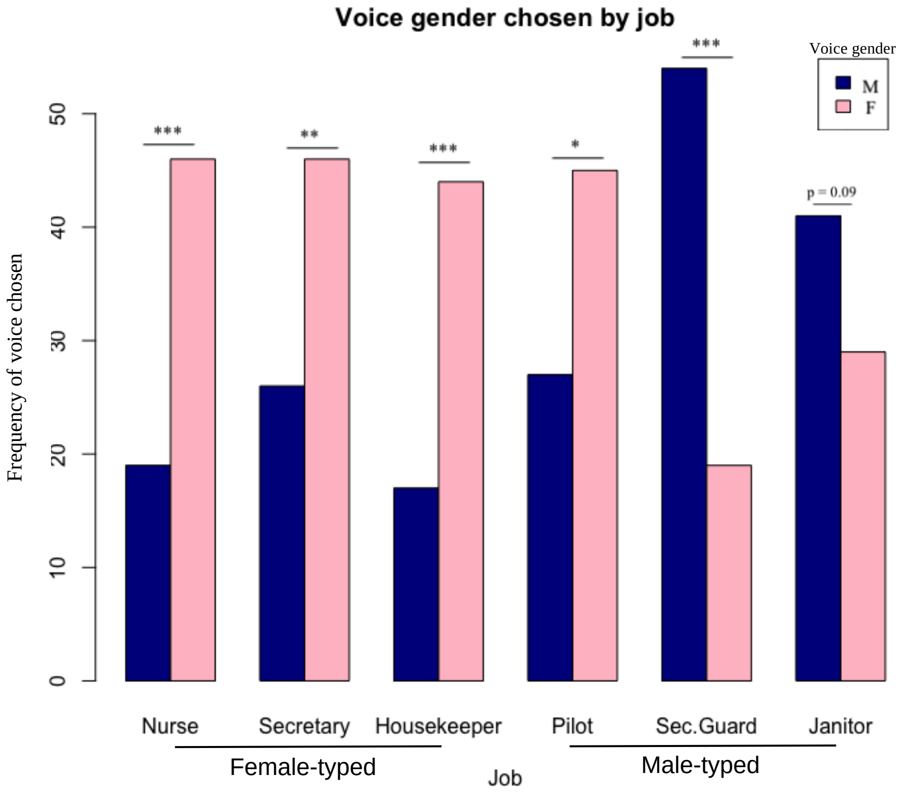


Figure 3. Stimuli from the Experimental Survey of Riley the Fake AI Baking Agent

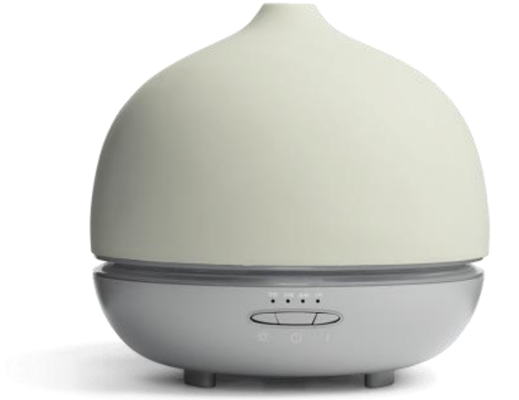
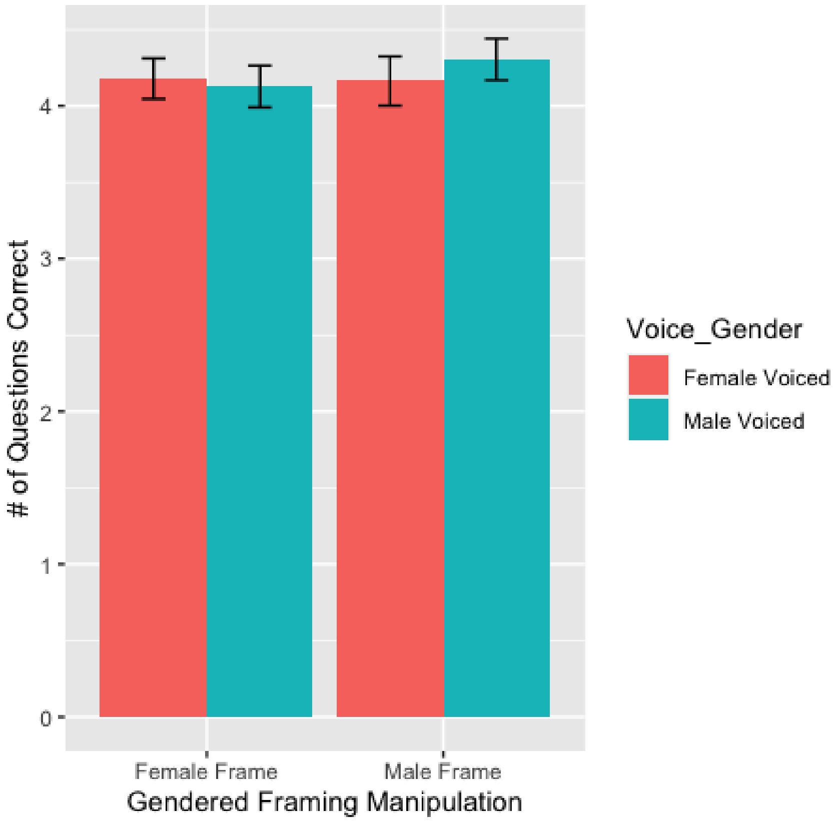


Figure 4. Stereotype Congruence on Actual Understanding



Error bars show 95% CI.

Figure 5. Stereotype Congruence on Perceived Understanding

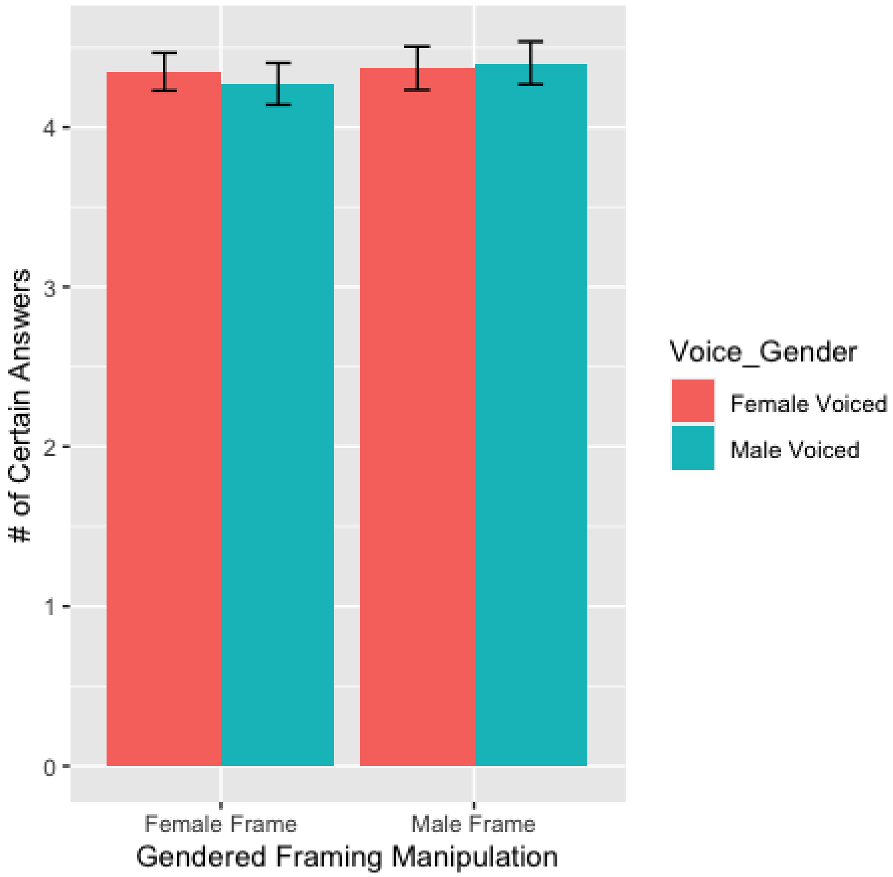
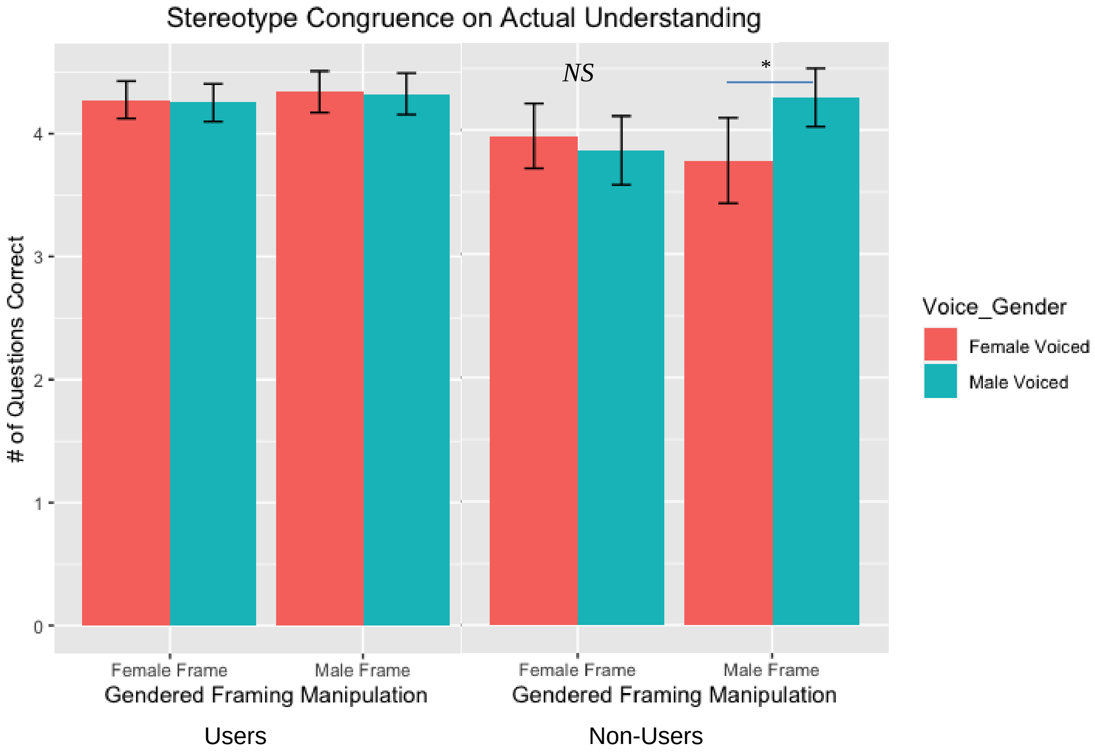


Figure 6. Comparing DVA Users vs. Non-Users



DVA users (4 left most columns) compared to non-users (4 right most columns)

Tables

Table 1. OLS Regressions Predicting Actual Understanding of Digital Voiced Assistant Dependent on Stereotype Congruence of Features

	Model 1	**Model 2**	**Model 3**	**Model 4**
Role Gender Male Frame	-0.027 (0.104)	-0.036 (0.105)	0.055 (0.123)	0.063 (0.124)
Voice Gender Male V	-0.056 (0.101)	-0.054 (0.102)	-0.022 (0.119)	-0.015 (0.120)
Status	0.092 (0.056)	0.098† (0.057)	0.077 (0.055)	0.082 (0.057)
Role Gender Male Frame × Voice Gender Male V	0.192 (0.144)	0.188 (0.144)	0.005 (0.170)	-0.030 (0.171)
Age		0.001 (0.003)		0.002 (0.003)
SS Gender *(Reference Group: Male)*				
Female		-0.017 (0.074)		-0.017 (0.073)
Other Gender		0.509 (0.531)		0.660 (0.524)
Employment Status *(Reference Group: Employed)*				
Unemp. Search		-0.349* (0.150)		-0.342* (0.148)
Unemp. Not search		-0.199 (0.141)		-0.171 (0.139)
Retired		-0.129 (0.178)		-0.094 (0.176)
Student		0.054 (0.235)		0.076 (0.231)
Political Affiliation *(Reference Group: Republican)*				
Democrat		-0.050 (0.092)		-0.059 (0.091)
Independent		-0.096 (0.102)		-0.080 (0.100)
Other Party		-0.185 (0.287)		-0.174 (0.282)

	Model 1	**Model 2**	**Model 3**	**Model 4**
Used DVA?			-0.307† (0.159)	-0.272† (0.160)
Role Gender Male Frame × Used DVA?			-0.261 (0.223)	-0.320 (0.224)
Voice Gender Male V × Used DVA?			-0.115 (0.218)	-0.145 (0.220)
Role Gender Male Frame × Voice Gender Male V × Used DVA?			0.629* (0.306)	0.720* (0.309)
Constant	3.418*** (0.469)	3.453*** (0.503)	3.635*** (0.469)	3.609*** (0.503)
R^2	0.010	0.026	0.049	0.064
Adjusted R^2	0.004	0.004	0.036	0.036
F Statistic	1.579	1.161	3.931	2.286

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 2. OLS Regressions Predicting Certainty of Understanding of Digital Voiced Assistants Dependent on Stereotype Congruence of Features for All Users

	Model 1	Model 2	Model 3	Model 4
Role Gender Male Frame	0.012 (0.097)	-0.006 (0.097)	0.044 (0.114)	0.035 (0.114)
Voice Gender Male V	-0.081 (0.094)	-0.073 (0.094)	-0.052 (0.110)	-0.044 (0.111)
Status	0.088† (0.052)	0.092† (0.053)	0.076 (0.051)	0.080 (0.052)
Role Gender Male Frame × Voice Gender Male V	0.108 (0.133)	0.108 (0.133)	-0.038 (0.157)	-0.058 (0.158)
Age		-0.002 (0.003)		-0.001 (0.003)
SS Gender *(Reference Group: Male)*				
Female		-0.075 (0.069)		-0.077 (0.068)
Other Gender		0.312 (0.491)		0.420 (0.485)
Employment Status *(Reference Group: Employed)*				
Unemp. Search		-0.278* (0.139)		-0.267† (0.137)
Unemp. Not search		-0.187 (0.130)		-0.161 (0.1299)
Retired		0.059 (0.164)		0.093 (0.163)
Student		0.054 (0.217)		0.057 (0.214)
Political Affiliation *(Reference Group: Republican)*				
Democrat		-0.062 (0.085)		-0.070 (0.084)
Independent		-0.148 (0.094)		-0.134 (0.093)
Other Party		-0.202 (0.265)		-0.191 (0.262)
Used DVA?			-0.325* (0.147)	-0.301* (0.148)
Role Gender Male Frame × Used DVA?			-0.097 (0.207)	-0.131 (0.208)

	Model 1	**Model 2**	**Model 3**	**Model 4**
Voice Gender Male V × Used DVA?			-0.101 (0.202)	-0.114 (0.204)
Role Gender Male Frame × Voice Gender Male V × Used DVA?			0.490† (0.284)	0.549† (0.286)
Constant	3.625*** (0.434)	3.831*** (0.465)	3.816*** (0.434)	3.974*** (0.466)
R^2	0.008	0.027	0.044	0.061
Adjusted R^2	0.002	0.004	0.031	0.032
F Statistic	1.271	1.195	3.503	2.154

* p < 0.1, ** p < 0.05, *** p < 0.01

Table 3. OLS Regressions Predicting Self-Reported Perceived Understanding of Digital Voiced Assistant Dependent on Stereotype Congruence of Features For All Users

	Model 1	**Model 2**	**Model 3**	**Model 4**
Role Gender Male Frame	0.917 (2.369)	1.428 (2.371)	0.598 (2.817)	1.456 (2.818)
Voice Gender Male V	-1.911 (2.304)	-2.161 (2.302)	-2.568 (2.733)	-2.451 (2.731)
Status	2.543* (1.270)	2.185† (1.291)	2.258† (1.269)	1.838 (1.289)
Role Gender Male Frame × Voice Gender Male V	1.872 (3.261)	1.242 (3.257)	1.530 (3.893)	0.489 (3.897)
Age		0.182* (0.075)		0.2000** (0.075)
SS Gender *(Reference Group: Male)*				
Female		1.965 (1.680)		2.069 (1.670)
Other Gender		6.283 (12.008)		8.466 (11.950)
Employment Status *(Reference Group: Employed)*				
Unemp. Search		-1.272 (3.392)		-0.731 (3.377)
Unemp. Not search		1.668 (3.190)		2.457 (3.177)
Retired		-3.870 (4.021)		-2.712 (4.012)
Student		0.974 (5.313)		0.892 (5.280)
Political Affiliation *(Reference Group: Republican)*				
Democrat		-2.850 (2.086)		-3.135 (2.073)
Independent		-3.615 (2.299)		-3.355 (2.282)
Other Party		-10.917† (6.482)		-10.195 (6.443)
Used DVA?			-8.073* (3.737)	-7.631* (3.647)
Role Gender Male Frame × Used DVA?			1.230 (5.112)	0.085 (5.121)

	Model 1	**Model 2**	**Model 3**	**Model 4**
Voice Gender Male V × Used DVA?			2.144 (5.112)	0.646 (5.024)
Role Gender Male Frame × Voice Gender Male V × Used DVA?			1.466 (7.023)	2.832 (7.057)
Constant	52.907*** (10.644)	49.934*** (11.372)	57.697*** (10.751)	54.219*** (11.483)
R^2	0.008	0.027	0.044	0.061
Adjusted R^2	0.002	0.004	0.031	0.032
F Statistic	1.271	1.195	3.503	2.154

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 4. OLS Regressions Predicting Understanding of Digital Voiced Assistant Dependent on Stereotype Congruence of Features for Novice Users

	Model 1	**Model 2**	**Model 3**	**Model 4**	**Model 5**	**Model 6**
Role Gender Male Frame	-0.200	-0.283	-0.050	-0.094	1.902	1.802
	(0.205)	(0.204)	(0.196)	(0.196)	(4.802)	(4.793)
Voice Gender Male V	-0.120	-0.125	-0.144	-0.103	-0.211	-1.299
	(0.202)	(0.202)	(0.194)	(0.195)	(4.744)	(4.756)
Status	0.010	-0.029	0.040	0.003	1.401	-1.218
	(0.112)	(0.118)	(0.107)	(0.114)	(2.617)	(2.773)
Role Gender Male Frame ×	0.627*	0.690*	0.449†	0.456†	2.913	2.289
Voice Gender Male V	(0.280)	(0.280)	(0.268)	(0.270)	(6.571)	(6.591)
Age		0.005		-0.002		0.320*
		(0.006)		(0.006)		(0.140)
SS Gender *(Reference Group: Male)*						
Female		-0.089		-0.242†		-0.877
		(0.149)		(0.143)		(3.495)
Other Gender		0.962		0.436		3.259
		(0.724)		(0.696)		(17.003)
Employment Status *(Reference Group: Employed)*						
Unemp. Search		-0.292		-0.010		5.683
		(0.263)		(0.253)		(6.186)
Unemp. Not search		-0.377		-0.304		1.163
		(0.243)		(0.234)		(5.714)
Retired		-0.111		0.193		-5.082
		(0.268)		(0.258)		(6.301)
Student		0.966†		0.524		8.569
		(0.570)		(0.549)		(13.398)
Political Affiliation *(Reference Group: Republican)*						
Democrat		-0.440*		-0.382*		-10.969*
		(0.180)		(0.174)		(4.237)
Independent		-0.491*		-0.499**		-13.850**
		(0.191)		(0.183)		(4.478)

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Other Party		-0.810† (0.441)		-0.599 (0.424)		-13.598 (10.358)
Constant	3.875*** (0.922)	4.447*** (1.003)	3.791*** (0.883)	4.659*** (0.965)	56.615** (21.627)	74.575** (23.569)
R^2	0.042	0.129	0.028	0.107	0.010	0.096
Adjusted R^2	0.021	0.060	0.008	0.036	-0.011	0.024
F Statistic	2.023	1.865	1.360	1.501	0.462	1.338

* p < 0.1, ** p < 0.05, *** p < 0.01