

PREFERENCE FOR GENDER STEREOTYPICALITY IN ARTIFICIAL INTELLIGENCE

BY

JULIA SPIELMANN

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Psychology
in the Graduate College of the
University of Illinois Urbana-Champaign, 2022

Urbana, Illinois

Doctoral Committee:

Assistant Professor Chadly Stern, Chair and Director of Research
Professor Dov Cohen
Associate Professor Catharine Fairbairn
Associate Professor Nathan Todd
Clinical Assistant Professor Andrea Miller

ABSTRACT

Do people prefer for voice-based artificial intelligence (AI) to align with gender stereotypes?

Gender is a salient and ubiquitous organizing principle that provides a sense of structure and simplicity to society. Initial evidence suggests that non-human entities, such as weather patterns, numbers, robots, and even AI, are perceived within a human-gender context. However, no research to date has investigated whether people prefer gender stereotypicality in AI. AI provides a unique context to examine gender stereotyping that is removed from both social (e.g., gender discrimination laws that apply to humans) and non-social factors (e.g., biological influences on gender roles among humans). Across four studies using experimental designs, I examined whether people prefer voice-based AI that aligns with gender stereotypes. Across Studies 1 and 2, I found that people preferred gender stereotypicality (over counterstereotypicality and androgyny) in voice-based AI. In Studies 3 and 4, I found that people perceived AI as more credible when the gender of the AI voice was a stereotypical (versus counterstereotypical) match to the gender of the question, which in part explained people's preference for stereotypicality. These studies contribute to understanding whether people prefer gender stereotypicality outside human gender relations, which holds implications for how AI might be used to both create and reinforce a gendered social reality.

ACKNOWLEDGMENTS

First and foremost, I would like to express immense gratitude to the best advisor I could have asked for, Chadly Stern. I am incredibly thankful for his continued guidance and support throughout my graduate studies. His dedication and enthusiasm for research, and his ability to teach and mentor me through the scientific process were invaluable. His patient advice, constructive criticism, and unconditional encouragement pushed and motivated me to become the researcher I am today.

I would also like to thank my committee members for their constructive feedback, suggestions, and thought-provoking questions, which greatly improved the quality of this work.

A special thanks to Rebecca Ginsburg and all Education Justice Project members, inside and outside, for their care and support, and for providing space for critical reflection.

My deepest gratitude to Mama and Papa, Oma, and my aunts for their love, words of encouragement, and care packages and cards that always lifted my spirits – *Danke!* Thank you to my sister Nathalie for your love and support, and sending cute videos as encouragement.

I am incredibly grateful for my UIUC friends, who have made this time special for me. To Tony, thank you for always lifting me up, supporting me in all my endeavors, and for your unwavering belief in me. I am appreciative of all the time we spent together in the “incubator”, your coding magic, and that I can always count on your help. Thank you for making me laugh like no one else, countless late-night grocery trips and Friday night extravaganzas, silly dances in our tiny kitchen, and your hugs. I am so grateful to have done this with you by my side. To Randi, I could not have wished for a better friend throughout this journey. Your support, encouragement, and genuine care have been invaluable. I am immensely grateful for our many mini-adventures, our pizza/ice cream/cookie baking traditions, our hot takes, our many

catchphrases, our need to turn simple activities into exciting challenges, and am coming to terms with all the times you made me sit through the end of a show. To Salvador for providing critical feedback and support. To Yarden for countless tea/coffee talks and delicious dinners, listening to me talk through my ideas, and just being an amazing supportive friend. To Miriam and Hany for wonderful and laughter-filled movie and food nights. To W.S., Big Chris, and friends for always supporting me.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION	1
CHAPTER 2: PILOT STUDIES	13
CHAPTER 3: PREFERENCE FOR GENDER STEREOTYPICALITY (STUDIES 1 & 2)	45
CHAPTER 4: MECHANISM DRIVING PREFERENCE FOR GENDER STEREOTYPICALITY (STUDIES 3 & 4).....	65
CHAPTER 5: GENERAL DISCUSSION	91
REFERENCES	103

CHAPTER 1: INTRODUCTION

Electronic technology, algorithms, and artificial intelligence (AI) have become an integral part of people's lives. As of January 2019, over 100 million Alexa AI devices have been sold worldwide (Bohn, 2019) and over 41.4 million people actively use Siri monthly in the US (Perez, 2017). Apple's feminine-voiced Siri and Amazon's feminine-voiced Alexa are voice-assistants developed to carry out tasks, such as checking the weather or creating a shopping list (Fessler, 2017; Purtill, 2021). In contrast, IBM's masculine-voiced AI named Watson is a question-answering computer system and is currently used to manage decisions in cancer treatment (Upbin, 2013).

The "gender" of AI is not only projected through the voice, but also through its name and description. For instance, when asked its gender, Alexa describes *itself* as "female in character", whereas IBM's Watson is referred to as "he" (Bort, 2015). Similarly, Microsoft's Cortana stands as another example of gender-based technology, whose name is in reference to a nude feminine video game character (Webster, 2019). Examples of voice-based AI fulfilling stereotypically feminine or masculine tasks suggest that people use gender as a salient organizing principle to make sense of AI. However, whether people prefer for voice-based AI to align with gender stereotypes (i.e., prefer *gender stereotypicality* in AI) has not yet been experimentally tested.

The gendered voices and names of these AI examples suggest that AI is programmed, perceived, and applied within a human-gender context. Artificial intelligence is a subfield of computer science that describes the theory, design, and study of computational agents simulating human intelligence (Legg & Hutter, 2007; Nilsson, 1980; Poole & Mackworth, 2010; Russell & Norvig, 2016). An intelligent agent is capable of learning, thinking, and responding in a similar manner as a human within a given social environment. It performs tasks that usually would

necessitate human intelligence, such as natural language processing, translation of languages, speech recognition, and planning. AI provides a unique context removed from both social (e.g., gender discrimination laws that apply to humans) and non-social factors (e.g., biological influences on gender roles among humans) to examine gender stereotyping. Any observed preferences for gender stereotypicality of AI would primarily exist due to cultural stereotypes, which holds implications for how AI might be used to both create and reinforce a gendered social reality.

Is a coherence between gender of AI voice and domain based on a systematic preference for gender stereotypicality? A recent UNESCO report (West et al., 2019) suggests that the answer to this question might be yes. The report highlights that gender-based inequality is pervasive in science, technology, engineering, and mathematics (STEM) fields and that societal gender biases influence and permeate the construction of technology. Further, UNESCO calls attention to the lack of knowledge on the impact of gendered technology on society. The present research that I outline here is the first of its kind to experimentally test people's preference for gender stereotypicality in AI and an associated mechanism driving people's preference. Specifically, consequences of and mechanisms through which a disembodied piece of coded technology "acquires" human gender and is treated in accordance with human gender stereotypes are yet unexplored.

Across four studies, I addressed two primary questions. First, I experimentally investigated whether people prefer congruence between gender of AI voice and gender domain of a question (i.e., preference for gender stereotypicality in AI). Second, I examined a potential mechanism that could contribute to explaining people's preference for gender stereotypicality in AI. Specifically, I examined whether perceiving an AI in a gender stereotypical domain as more

credible would in part explain the preferred relation between gender domain of a question and gender of AI voice.

Prevalence of Gender Stereotypes

Anecdotal evidence highlights the prominence of gendered AI, which prompts reason to believe that the application of human gender stereotypes to technology is purposeful and systematic in nature. Researchers have long drawn a distinction between gender and sex (Bem, 1993; Eagly & Wood, 2016; Wood & Eagly, 2015). Gender refers to the socially and culturally shared meaning ascribed to male and female social categories (Wood & Eagly, 2015). Men and women as members of a gender category differ in their perceived and enacted masculinity and femininity. Roles associated with gender are socially constructed and performative (Butler, 2004; Deaux & Major, 1987; Morgenroth & Ryan, 2018, 2020). Multiply determined gender roles are continuously reinforced throughout history and across cultures (Bussey & Bandura, 1999; Spence & Helmreich, 1979). For instance, women are expected to be compassionate, empathic, warm, and helpful, and in turn are perceived as well-suited for childcare duties (Bussey & Bandura, 1999; Deaux & Major, 1987; Eagly & Mladinic, 1994).

In contrast to gender, sex differentiates people based on the anatomy of their reproductive organs and hormones and is usually assigned at birth (Wood & Eagly, 2015). Albeit that sex and gender roles often align, they do not inherently need to. For instance, a person who is biologically male may prefer to take on a stereotypically feminine role (e.g., nurse), whereas a person who is biologically female may have an interest in a masculine role (e.g., construction).

Why might gender be applied to AI? Gender is arguably one of the most salient and important organizing principles within social cognition (Banaji & Hardin, 1996; Diekmann & Schmader, 2021; Kurzban et al., 2011; Macrae & Bodenhausen, 2000; Martin & Mason, 2022).

Gender is ubiquitous in all cultures (Bem, 1981b, 1993; Eagly & Wood, 1999), and people readily use gender to ascribe social roles (Eagly & Wood, 2011). As a function of its primacy in social judgment, perceivers automatically and inevitably encode a person's gender (Kurzban et al., 2001). Thus, gender provides a sense of structure, simplicity, and predictability to society (Bem, 1981b, 1993; Eagly & Wood, 1999).

Schemas and stereotypes based on gender categories act as lenses and facilitate the processing of gender-relevant information (Bem, 1974, 1993). Reliance on gender as a heuristic allows people to navigate the world more efficiently and helps to uphold the division of men and women into social roles (Bem, 1981b; Martin & Halverson, 1983; Taylor et al., 1979). For instance, gender is one of the few social categories in which it is socially acceptable to segregate individuals (e.g., single-gender schools and classrooms) or provide separate spaces based on a person's group membership (e.g., bathrooms; Billger, 2009; Maccoby & Jacklin, 1987). As such, societal structures derived from gender are a central part of everyday life used to organize and make sense of the world. The salience and ubiquity of gender across cultures highlight that the application of gender potentially occurs beyond interaction with humans. Specifically, socially created aspects of human gender might also be applied to non-human entities.

Through a systematic application of gender to technology, human behavioral prescriptions and expectations based on gender likely extend to AI. As noted above, coded technology projects its gender through name, voice, or appearance. Beyond documenting that people gender technology, however, little is known about whether people prefer gender stereotypicality in technology and which mechanism drives this preference.

Research examining gender stereotype content about humans suggests that people's consensual beliefs about men and women are closely tied to desired personality traits and

behavioral expectations (Abele, 2003; Eagly, 1987; Fiske et al., 2002; Fiske et al., 2007). Gender stereotypes are shared beliefs about how certain attributes differentially apply to men and women (e.g., Eagly 1987; Fiske & Stevens, 1993). Social role theory posits that social structures, such as differences in social roles among men and women, underlie and account for gender stereotypes (Eagly & Steffen, 1984; Eagly, 1987). Research highlights that repeated association of a particular social role with men or women results in greater gender stereotypicality of that domain. For example, men and women implicitly and explicitly associate masculinity with engineering, science, math, career, and leadership, and associate femininity with teaching, humanities, and family (Koenig et al., 2011; Nosek et al., 2002; Nosek et al., 2007; White & White, 2006).

A wealth of research has documented a strong preference for congruence between gender and social roles (e.g., role congruity theory; Diekmann & Goodfried, 2006; Diekmann & Hirnisey, 2007; Eagly & Karau, 2002; Jackson, 1983; Koch et al., 2015). I utilize the term “preference for gender stereotypicality” to describe the preference for consistency between demands of a social role and characteristics of its occupants. In other words, a strong preference for gender stereotypicality implies that people positively evaluate when characteristics of a group are in alignment with the group’s prototypical social roles. I argue that common knowledge of gender stereotypes acquired through role observation, engagement with others, and cultural exposure could subsequently lead to preferences for gender stereotypicality in non-human entities. Specifically, the content of gender stereotypes might create different preferred behaviors and roles for gendered AI.

Why AI?

What do we stand to theoretically gain from examining preferences for gender stereotypicality in *artificial intelligence*? I argue that AI represents a unique space to examine questions of human cognitive processes for two primary reasons. First, AI is a medium removed from *non-social factors*, such as biological influences of gender roles. AI offers an inherent “blank slate”, as it is a disembodied agent without reproductive organs, biological sex, hormones, or emotions. Thus, any preference for gender stereotypicality in AI is constructed entirely by humans. One study from Hoffman and Hurst (1990) provides initial support for this reasoning. The researchers examined gender stereotypical association between traits and categories in a *hypothetical space* (i.e., one that began as a blank slate). Participants read descriptions of fictitious groups with unequal divisions of labor and made trait judgments about members of these groups. The researchers found that participants formed role-based stereotypes, which affected trait evaluations even when a group member’s role was not specified. These findings suggest that gender stereotypes arise and acquire their domain specific content to explain (and possibly justify) differences in status and social roles even outside of the prevailing societal gendered context. In other words, fictitious groups served to capture the rationalization and justification of gender distribution into different social roles. Although Hoffman and Hurst’s (1990) methodology allowed for a close investigation of gender stereotype development in a space removed from social norms, the hypothetical space still involved individuals with human-like traits. AI represents a cleaner “blank slate” likely used to both create and reinforce social reality. Importantly, human interaction with AI has the potential to provide novel insight into preference for gender stereotypicality outside of current societal gender relations.

Second, AI is a medium removed from *social factors*, such as gender discrimination laws that apply to humans. AI is uniquely situated to examine stereotypes as it creates a space distant from social norms surrounding the expression of stereotypes and negative attitudes. Contemporary social psychology research has documented a decrease in the expression of explicit negative attitudes toward members of marginalized groups (e.g., Black people; Dovidio & Gaertner, 1986; Dovidio et al., 1998). However, changes in self-reported negative attitudes over time do not necessarily reflect an actual change in attitudes. Instead, the change might result from increased social desirability norms surrounding the explicit expression of biases (Hoffmann & Musch, 2019). While a decrease in explicit biases is normatively beneficial, self-presentation strategies mask the extent to which people endorse negative attitudes and biases. Thus, psychological processes that continue to occur in people's minds remain largely unclear.

Examining social processes in AI affords the unique opportunity to test for gender biases while minimizing the influence of social desirability. Disinhibition and violation of conversational norms follow from a lack of retaliation by the non-human entity. For instance, a lack of pushback from an AI likely allows people to express biases in unsupervised interactions (e.g., calling an AI a misogynistic slur would not result in pushback from the AI). Social norms surrounding correct use of language, behavioral inhibition, abuse, and misuse of intelligent agents seem to be perceived as less strong (Bartneck et al., 2005; Brahnam & De Angeli, 2008; De Angeli, 2009; Sanchez Ramos et al., 2018). For example, Sanchez Ramos and colleagues (2018) found that people are more likely to engage in verbal abuse of humanoid robots than of humans. Non-human agents that are attributed less mind are also more frequently verbally abused (Keijsers & Bartneck, 2018). Taken together, these findings suggest that people have little to no fear of retaliation from non-human entities when violating conversational norms and

engaging in socially stigmatized behavior with a non-human entity. Thus, usually inhibited gender biases likely are expressed more freely when interacting with AI. Overall, the “blank slate” and greater potential for uncensored expression of gender biases uniquely positions AI to investigation of preferences for gender stereotypicality.

Making Sense of Non-Human Entities through Application of Human Gender Stereotypes

Is technology a means through which societal structures of gender are expressed (Bray, 2012; Robertson, 2010)? Some feminist theorists initially posited that computers and robots would eradicate the use of human gender (Haraway, 1987, 1997). Specifically, Haraway rejected gender essentialism and argued that technology affords the possibility to construct one's (gender) identity outside of existing social categorization. The Cyborg Manifesto (Haraway, 1985) inspired hope that change in societal structures could occur through technology. Ideas about disrupting binary thinking are adjacent to the intergroup contact hypothesis (Allport, 1954), which initially motivated hope that under certain conditions intergroup contact could successfully reduce group-based negative attitudes. However, findings demonstrate that the positive effects of intergroup contact on reduction of prejudice are small and inconsistent (Schaller & Crandall, 2003; Gaertner et al., 2000; Pettigrew & Tropp, 2008). This may especially hold true for gender relations, as men in heterosexual relationships often continue to hold sexist beliefs that legitimize gender inequality (Glick & Fiske, 2001). Further, despite changes in gender roles in the last decades, gender stereotypes continue to persist (Diekmann & Eagly, 2000; Eagly et al., 2020). Basic cognitive functions, such as social categorization and associating certain expectations with a category, are not easily disrupted. Taken together, this evidence suggests that categorical thinking and preferences for how men and women ought to behave likely persist through gendered technology.

In contrast to Haraway's initial theorizing, some empirical evidence suggests that the automatic application of social categories extends even to non-human entities (Nass & Moon, 2000; Nass et al., 1997). Similar to human categorization, people automatically categorize robots based on their perceived race, gender, and nationality (Carpenter et al., 2009). Further, masculinity and femininity are applied to non-human objects, such as toys or clothing (Auster & Mansbach, 2012; Beasley & Collins Standley, 2002; Weisgram et al., 2014), and to abstract non-social concepts, such as shapes, weather patterns, or Arabic numbers (Boroditsky et al., 2003; Jung et al., 2014; Stroessner et al., 2020; Wilkie & Bodenhausen, 2012, 2015). For instance, Wilkie and Bodenhausen (2012, 2015) found that odd numbers were perceived as more masculine and even numbers as more feminine. Gendered names given to hurricanes are also associated with biased risk perception (Jung et al., 2014). Gender-based expectations impact people's preparedness, resulting in more hurricane related deaths for female named hurricanes than male named hurricanes. Thus, gender categorization extends to non-human constructs and entities.

Emerging research examining the use of social categories and stereotypes to make sense of non-human entities is still in its infancy. In interactions with non-human entities, gender is used as an important organizing principle to make sense of available information. However, little is known about whether people's tendency to categorize and stereotype non-human entities translates into a specific preference for gender stereotypicality in AI. Indeed, to the extent of my knowledge, only one study has previously examined this. Specifically, one study suggests that people categorize and subsequently differentially evaluate gendered robots. The authors observed that people were more likely to choose to work on a stereotypically masculine task (i.e., math) compared to a stereotypically feminine task (i.e., verbal) with a robot with masculine short

hair, whereas people were equally likely to choose to work on a masculine or feminine task with a robot with feminine long hair (Eyssel & Hegel, 2012). This finding suggests that a robot's gendered facial cues, such as hair length, bias perception and result in preference for stereotypicality. Although this study provides some initial evidence for gender stereotypicality preference, it is limited in scope due to the small sample size ($N = 60$) and focuses on humanoid robots instead of non-human entities without physicality. Further, the study cannot speak to mechanisms of a preference for stereotypicality in AI, which requires further investigation.

Overall, the literature on gendering of machines documents that people gender non-human entities. However, the extent to which people prefer gender stereotypicality in AI and what psychological mechanism is associated with this preference has not yet been examined. In the present research, I examine preferences for gender stereotypicality in AI. In the *preference for gender stereotypicality hypothesis*, I predicted that people would prefer a feminine-voiced (over masculine-voiced) AI more strongly when asked to answer a question in a stereotypically feminine (versus masculine) domain.

Participant Sex/Gender as a Potential Moderator

A wealth of research consistently demonstrates that both men and women use gender stereotypes and penalize people who deviate from them (Heilman et al., 2004; Rudman & Glick, 1999, 2001; Rudman & Phelan, 2008; Rudman et al., 2012). However, a novel context removed from societal norms surrounding the expression of negative attitudes, such as AI, might present an exception and allow for gender differences in the treatment of AI to emerge. Interestingly, some evidence suggests that men direct more backlash in response to counterstereotypicality than women (Carli & Eagly, 2001; Prentice & Carranza, 2004), and counterstereotypical behavior in men (compared to women) is evaluated more negatively and penalized more harshly

(McCreary, 1994; Sirin et al. 2004). In the context of AI, these differences might become even more pronounced. Examining gender differences in preferences for stereotypicality in AI might shed light on whether men in particular regulate their biases more strongly than women in everyday life. Men might feel more comfortable expressing preference for gender stereotypicality and endorsing gender stereotypes in the context of AI due to the lack of potential repercussions. This line of reasoning highlights the need for exploration of gender differences in the present work. In addition, prior research has not yet disentangled the assessment of sex from gender and tested whether participant sex and gender differentially modulate relations. In studies 1 and 2, I examined moderation by participant sex, and in studies 3 and 4, I examined moderation by participant gender. This approach allowed me to examine whether effects differ across these different measurements.

The Present Research

In the present research, I examined people's preference for gender stereotypicality in AI and one contributing mechanism of this preference. In Study 1, I experimentally tested whether people would prefer gender stereotypicality (over counterstereotypicality) in voice-based AI. Specifically, I predicted that participants would show an increased preference for a feminine-voiced (over masculine-voiced) AI when asked to answer a question in a stereotypically feminine (versus masculine) domain. In Study 2, I examined how the option of an androgynous-voiced AI (i.e., a blend of prototypically feminine- and masculine-voices) would affect people's preferences. I outline competing predictions about people's preference for stereotypicality or androgyny below. In Study 3, I examined whether people perceive an AI as more credible when the gender of the AI voice is a stereotypical (versus counterstereotypical) match to the gender of the question. In Study 4, I examined whether perceiving an AI in a gender stereotypical domain

as more credible would in part explain the preferred relation between gender of question domain and AI.

CHAPTER 2: PILOT STUDIES

Pilot Study 1

I first conducted a pilot study to select stimuli sentences that are perceived as non-gender-typed (i.e., not implying a gender domain). Selected sentences are used as stimuli in subsequent studies.

Methods

Participants

328 participants (174 female participants, 152 male participants, 2 “Other” sex; $M_{\text{age}} = 42.23$; range = 20-77) were recruited from the Mechanical Turk (Mturk) website using the CloudResearch Approved Participant List, which only allows MTurk workers who have previously passed attention and engagement checks to complete a study (Litman et al., 2017). Three additional participants completed the study, but were excluded from analyses for failing an attention check. I aimed to collect enough participants to have at least 100 participants rate each stimulus while expecting around 10% of participants to fail the attention check. In turn, I needed to collect at least 300 participants for analyses, and so aimed to collect 330 participants. A sensitivity power analysis indicated that the final sample size provided 80% power to detect an effect size as small as $r = .15$. This power analysis was conducted using G Power 3.1 (Faul et al., 2007).

Materials

Non-gender-typed sentences. Forty-five stimuli sentences were created by the researcher and tested in this pilot study. Adjectives and nouns used to construct the sentences were taken from lists of words rated as non-gender-typed in previous research (Crawford et al., 2004; Scott et al., 2019). A list of all sentences can be found in Table 1.

Procedure

In all studies, participants were provided with an external link on the Mechanical Turk website (Buhrmester et al., 2011, 2018) that took them to the study. All studies were programmed using Qualtrics online-survey software. After providing consent, participants performed the tasks described below. At the end, participants completed an attention check and were debriefed.

Measures

Sentence rating task. In the first part of the study, participants read sentences presented in randomized order. Participants read the following instructions: “In this study we are interested in your perceptions of the world. On the following pages, you will be asked to rate sentences about different topics. You will be asked to indicate how feminine or masculine you perceive the sentences to be. A sentence is feminine if it is stereotypically more likely to be said by a woman than a man. A sentence is masculine if it is stereotypically more likely to be said by a man than a woman. A sentence is neutral if it is not stereotypically more likely to be said by a man or a woman.” Participants were randomly presented with 15 out of the 45 sentences and rated the sentences according to perceived masculinity-femininity on a 7-point scale (1 = *Extremely masculine*, 4 = *Neutral*, 7 = *Extremely feminine*; Crawford et al., 2004).

Demographics. Participants reported their age, race, sex, and political ideology. Participants’ sex was assessed using the following question: “What is your sex?” (*Male, Female, Other*: [open ended], *Prefer not to answer*).

Political ideology. Participants reported their political ideology in response a single item: “Where on the following scale of political orientation would you place yourself?” (1 =

Extremely liberal; 5 = *Moderate*; 9 = *Extremely conservative*). Previous research has found that this single item exhibits high levels of predictive validity (Jost, 2006).

Attention check. The attention check was presented to participants at the end of the survey. Participants read the following: “People vary in the amount they pay attention to these kinds of surveys. Some take them seriously and read each question, whereas others go very quickly and barely read the questions at all. If you have read this question carefully, please write the word yes in the blank box below labeled other. There is no need for you to respond to the scale below.” Below the information was a 7-point scale and response box.

Results

Non-gender-typed sentences. To select sentences for use in future studies, I used a two-step selection process. First, I used one-sample *t*-tests to compare masculinity/femininity ratings of each sentence against the mid-point of the scale (i.e., 4). Sentences with ratings that (1) had the smallest effect size, (2) did not significantly differ from mid-point of scale, and (3) had the smallest standard deviation were selected as non-gender typed sentence stimuli for future studies. I selected sentences 1-19 using the above criteria. Second, I also examined whether ratings for sentences 1-19 correlated with participant sex (effect coded as female = -1, male = 1) and political ideology. Participants who selected having an “other” sex were excluded from these analyses. Given that participant sex and/or political ideology might be variables of interest in examining preference for gender stereotypicality in future studies, I sought to select sentences with the smallest correlations between gender stereotypicality ratings and participant sex and political ideology. Based on these criteria, I selected sentences 1-14 (see Table 1).

Table 1

Means, standard deviations, one sample t-tests for gender stereotypicality ratings, and correlations of stereotypicality ratings with participant sex and political ideology for Pilot Study 1 sentences

Note. The first 14 sentences were used in Pilot Study 2. Bolded sentences were used in all studies.

* $p < .05$; ** $p < .01$

#	Sentence	Gender Stereotypicality Rating								Correlation with Stereotypicality Ratings	
		<i>M</i>	<i>SD</i>	<i>t</i>	df	<i>p</i> -value	<i>M</i> _{diff}	CI-L	CI-U	Participant Sex	Political Ideology
1	The water is deep.	4.09	1.00	0.09	111	.343	0.09	-0.10	0.27	-.03	-.02
2	Trees make up a forest.	3.98	0.76	0.07	105	.798	-0.02	-0.16	0.13	.01	.07
3	The bus stopped.	3.94	0.58	0.06	106	.320	-0.06	-0.17	0.06	.12	-.03
4	I need confirmation.	4.02	1.29	0.12	108	.882	0.02	-0.23	0.26	-.09	.11
5	I am a client	3.98	0.89	0.09	107	.828	-0.02	-0.19	0.15	.04	.07

Table 1 (cont.)

6	I have a bookcase.	3.98	0.76	0.07	110	.804	-0.02	-0.16	0.13	.05	.01
7	This is an address.	4.02	0.67	0.06	111	.779	0.02	-0.11	0.14	.14	.14
8	This situation is thrilling.	4.05	1.24	0.12	108	.701	0.05	-0.19	0.28	-.10	.03
9	I like coffee.	4.05	0.79	0.08	107	.544	0.05	-0.10	0.20	.08	-.07
10	This is great.	4.08	0.96	0.09	108	.373	0.08	-0.10	0.27	.02	.04
11	It's 3 o'clock.	3.94	0.56	0.05	107	.306	-0.06	-0.16	0.05	-.19	.11
12	This movie is great.	4.12	1.05	0.10	110	.243	0.12	-0.08	0.31	-.03	.11
13	Water is tasteless.	3.87	1.10	0.10	114	.205	-0.13	-0.33	0.07	-.04	-.07
14	The tree is green.	4.11	0.72	0.07	108	.116	0.11	-0.03	0.25	-.10	.03

Table 1 (cont.)

15	Horses are animals.	3.99	0.80	0.08	108	.905	-0.01	-0.16	0.14	.08	.21*
16	This is endless.	4.03	1.12	0.11	107	.796	0.03	-0.18	0.24	-.02	.29**
17	I lost my keys.	4.04	1.02	0.10	112	.646	0.04	-0.15	0.23	.04	.22*
18	Hamsters are animals.	4.11	0.88	0.08	108	.192	0.11	-0.06	0.28	-.76	.20*
19	I have change.	4.15	1.04	0.10	109	.145	0.15	-0.05	0.34	.21*	.10
20	The couch is next to the bookcase.	4.14	0.81	0.08	109	.079	0.14	-0.02	0.29	-	-
21	This is simple.	3.80	1.10	0.11	106	.068	-0.20	-0.41	0.02	-	-
22	A dog is playing with a ball.	3.84	0.90	0.90	109	.060	-0.16	-0.33	0.01	-	-
23	This is a caterpillar.	4.17	0.77	0.08	104	.024	0.17	0.02	0.32	-	-

Table 1 (cont.)

24	I eat macaroni for dinner.	3.74	1.04	0.10	108	.011	-0.26	-0.45	-0.06	-	-
25	Apples and grapefruits are fruits.	4.23	0.95	0.09	110	.011	0.23	0.05	0.41	-	-
26	I need medicine.	4.26	0.99	0.10	104	.009	0.26	0.07	0.45	-	-
27	There are fish in the water.	3.75	0.95	0.09	109	.008	-0.25	-0.42	-0.07	-	-
28	It's windy outside.	4.24	0.89	0.09	106	.006	0.24	0.07	0.41	-	-
29	It is evening.	4.25	0.90	0.09	105	.004	0.26	0.08	0.43	-	-
30	The water is warm.	4.24	0.79	0.08	108	.002	0.24	0.09	0.39	-	-
31	The road is uneven.	3.61	1.18	0.11	106	.001	-0.39	-0.62	-0.17	-	-
32	The airplane left the ground.	3.69	0.93	0.09	109	.001	-0.31	-0.48	-0.13	-	-

Table 1 (cont.)

33	The car goes fast.	3.43	1.24	0.12	114	<.001	-0.57	-0.79	-0.34	-	-
34	It's sunny outside.	4.36	0.87	0.08	108	<.001	0.36	0.19	0.52	-	-
35	The child is playing in the garden.	4.77	1.02	0.10	105	<.001	0.77	0.58	0.97	-	-
36	A carpenter built this house.	3.24	1.15	0.11	108	<.001	-0.76	-0.98	-0.54	-	-
37	The bread is freshly baked.	4.99	1.03	0.10	111	<.001	0.99	0.80	1.18	-	-
38	I want to borrow your book.	4.35	0.96	0.09	103	<.001	0.35	0.16	0.53	-	-
39	I am uncomfortable.	4.94	0.95	0.09	106	<.001	0.94	0.76	1.13	-	-
40	This robe is simple.	4.68	1.16	0.11	111	<.001	0.68	0.46	0.90	-	-
41	This table is great.	4.45	1.20	0.11	110	<.001	0.45	0.22	0.68	-	-

Table 1 (cont.)

42	This situation is silly.	4.85	1.25	0.12	108	<.001	0.85	0.62	1.09	-	-
43	It's warm on the beach.	4.72	0.95	0.09	111	<.001	0.72	0.55	0.90	-	-
44	I like to eat fruit.	4.56	0.91	0.09	110	<.001	0.56	0.39	0.73	-	-
45	I like to eat apples.	4.28	0.75	0.07	113	<.001	0.28	0.14	0.42	-	-

Note. CI-L and CI-U = Lower and Upper Bound of the 95% confidence interval around *t*-value, respectively.

Pilot Study 1 Summary

In this initial pilot study, I selected 14 non-gender-typed sentences for use in studies.¹

Pilot Study 2

I next conducted a second pilot study to select voices that (1) were uniformly perceived as a man's, woman's, or an androgynous voice (gender ratings), (2) were rated prototypical of the average man or woman (gender-typed voices) or were *not* rated as prototypical of the average man or woman (androgynous voices), and (3) were highest in perceived masculinity, femininity, and androgyny. I also tested stimuli questions to select stimuli that are perceived as stereotypically masculine or feminine.

Participants

323 participants (172 female participants, 148 male participants, 1 “Other” sex, 2 “Prefer not to answer”; $M_{\text{age}} = 42.36$; range = 18-80) were recruited from Mturk using the CloudResearch Approved Participant List. Eight additional participants completed the study, but were excluded from analyses for failing an attention check. I aimed to collect enough participants to have at least 100 participants rate each stimulus while expecting around 10% of participants to potentially fail the attention check. I needed to collect at least 300 participants for analyses, and so aimed to collect 330 participants. A sensitivity power analysis using G Power 3.1 indicated that the final sample size provided 80% power to detect an effect size as small as $r = .15$.

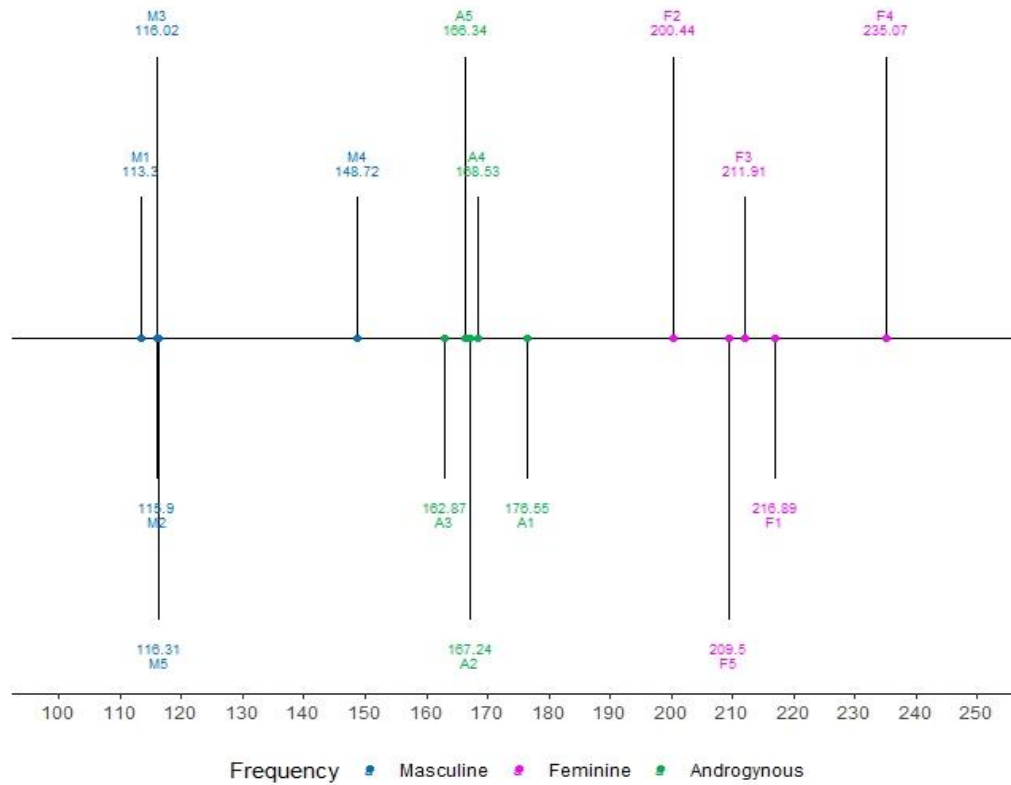
¹ I also pilot tested gender question domains in Pilot Study 1. However, given issues with measurement, I ultimately did not use these results. Instead, I amended the measure and assessed ratings in Pilot Study 2.

Materials

Vocal stimuli. Voices of five male and five female volunteers were recorded. Volunteers were healthy non-smoking cisgender White Midwesterners in their 20s and varied in the degree of gender stereotypicality of their voice perceived by the researcher. Volunteers recorded themselves stating the 14 non-gender-typed stimuli sentences selected in Pilot Study 1 in a non-emotional and neutral tone. Voices of ten volunteers were recorded to allow for a range of vocal stereotypicality. Within the human population, the average female voice pitch ranges from 170-235 Hz and the average male voice pitch ranges from 100-150 Hz (Simpson, 2009). Voices from female volunteers fell within the population range of female voices and were considered feminine, whereas voices from male volunteers fell within the population range of male voices and were considered masculine (see Figure 1). I used version 2.4.2. of Audacity(R) recording and editing software (Audacity Team, 2020) and version 6.1.51 of praat (Boersma & Weenink, 2020) to create five androgynous voices by changing the fundamental frequency (pitch) of the recorded feminine voices. Androgynous voice pitch (i.e., a more equal blend of masculine and feminine pitch) ranges from 140-180 Hz and are perceived as gender ambiguous (Dacakis, 2000; Junger et al., 2013; Simpson, 2009). All vocal stimuli were processed in praat and altered to approximately the same duration and intensity (i.e., 65 decibel).

Figure 1

Average pitch of each voice across sentences 1-14 tested in Pilot Study 2.



Gender domain specific questions. Sixty questions were constructed by the researcher and were intended to be difficult and have a single answer. Stimuli questions pertained to ten masculine domains (i.e., sports, mathematics, cars, hunting, fishing, wood working, astronomy, physics, architecture, aviation) and ten feminine domains (i.e., arts & crafts, beauty, childcare, librarian, dance, fashion, nutrition, emotion, teaching, cleaning). Domains were selected from previous research suggesting that they would be perceived as gendered (Misersky et al., 2014). Three questions per domain were tested (e.g., three questions about sports). A complete list of

piloted feminine domain questions can be found in Table 5 and a complete list of piloted masculine domain questions can be found in Table 6.

Procedure

After providing consent, participants read the following instructions: “At this time, please put your headphones on and turn the volume up on your device. You will be listening to sounds on the following pages.” Participants completed a sound check, and then performed the tasks described in the next section. Participants provided basic demographic information, such as their age, race, and gender. At the end, participants completed an attention check and were debriefed.

Measures

Sound check. To ensure that participants were listening to the vocal stimuli, I created a sound check. Participants were shown the following: “This is a sound check. Please have your headphones on and your volume turned up. Please listen to this audio.” Participants were automatically played the sound of a flowing river and asked what sound they heard (“A dog barking”, “Flowing water”, “Sirens”, “A beep”). If participants chose the correct answer (i.e., “Flowing water”), they were able to proceed to the main task. If participants failed to provide the correct answer, they were given one more chance to complete the sound check. If participants failed the sound check twice, they were redirected to the end of the survey.

Voice categorization and rating task. At the beginning of the task participants read the following instructions: “In this study, we are interested in your perceptions of the world. On the following pages, you will listen to short recordings of people's voices and be asked several questions. First, you will listen to short recordings of someone speaking. The recordings will start to play automatically. You can replay the recordings as many times as needed. Then, you will be asked to rate whether a man or a woman spoke. If you are unsure who spoke in the

recording, select the option "I don't know". Lastly, you will be asked to rate the masculinity and femininity of the voice. A voice is feminine if it is stereotypically more likely to be a woman's voice. A voice is masculine if it is stereotypically more likely to be a man's voice. A voice is neutral if it is not stereotypically more likely to be a woman's or a man's voice."

During the task participants were randomly assigned to listen to two masculine, two feminine, and two androgynous voices stating different non-gender-typed sentences. Participants indicated the perceived gender of the voice by categorizing the voice as "man", "woman", and "I don't know". Participants were only allowed to proceed to the next page once they spent three seconds on the page to ensure that they would listen to the vocal stimuli, but were given as much time as needed to make their categorization. If participants categorized the voice as either man or woman, they also were asked to indicate "How typical is the person's voice of a [e.g., man]?" Participants rated perceived typicality of the voice on a 7-point scale (1 = *Not at all typical*, 7 = *Extremely typical*). All participants also indicated how gender stereotypical the voice was. Participants rated the perceived masculinity or femininity of the voice on a 7-point scale (1 = *Extremely masculine*, 4 = *Neutral*, 7 = *Extremely feminine*).

Question rating task. In the second part of the study, participants read masculine and feminine questions presented in randomized order. Participants read the following instructions: "On the following pages, you will be asked several questions. First, you will be asked to rate whether you think a man or a woman would be better at answering different questions. Then, you will be asked to rate how interesting and difficult to answer the questions are, and how curious you are to find out the answers. When providing responses, imagine needing to answer these questions without any help from other humans or technology." Participants were randomly presented with 20 out of 60 questions and asked to indicate "On average, who do you think

would be more likely to answer the question correctly?” (1 = *A man would be more likely*, 2 = *A woman would be more likely*, 3 = *A man and a woman would be equally likely*).

I also assessed the level of difficulty, level of interest, and level of curiosity in regards to the questions. Participants were asked “How difficult to answer is this question?” and responded using a 7-point scale (1 = *Extremely easy*, 4 = *Neither easy nor difficult*, 7 = *Extremely difficult*). Participants were asked “How interesting is this question?” and responded using a 7-point scale (1 = *Not at all interesting*, 7 = *Extremely interesting*). Lastly, participants were asked “How curious are you to find out the answer to this question?” and responded using a 7-point scale (1 = *Not at all curious*, 7 = *Extremely curious*).

Results

Analytic plan

Given that each voice was shown with sentences 1-14, I created a composite item that combined gender categorizations across the different sentences for each voice. Similarly, I combined all prototypicality ratings and, separately, all stereotypicality ratings across sentences for each voice.

Voice categorization and rating task

To select voices for future studies, I used a multi-step selection process (see Tables 2-4 for statistics). I examined (1) gender categorization, (2) prototypicality ratings, and (3) gender stereotypicality ratings, and selected voices based on all three variables as outlined below.

Gender categorization. First, I examined the perceived gender of the vocal stimuli. Using Chi-square tests, I examined whether the observed sample distribution for perceived gender significantly differed from the expected distribution based on chance (i.e., 33.33%) for each voice. To obtain gender-typed voices, I selected voices that were perceived as significantly

more representative of one category (i.e., man or woman). To obtain androgynous voices, I selected voices which had the highest percent of “unsure” categorizations.

Prototypicality ratings. Next, I examined prototypicality ratings. I used one-sample *t*-tests to compare prototypicality ratings against the mid-point of the scale (i.e., 4) for each vocal stimulus. To obtain gender-typed voices, I selected vocal stimuli as being prototypical of either men or women if they (1) had the biggest effect size, (2) significantly differed from the mid-point of the scale, and (3) had the smallest standard deviation. To obtain androgynous voices, I selected vocal stimuli that were not significantly perceived as prototypical of the average man or woman, and that did not significantly differ from the mid-point of the scale.

Gender stereotypicality ratings. Lastly, I examined gender stereotypicality ratings. I used one-sample *t*-tests to compare prototypicality ratings against the mid-point of the scale (i.e., 4) for each vocal stimulus. To obtain gender-typed voices, I selected vocal stimuli with ratings that (1) had the biggest effect size, (2) significantly differed from the mid-point of the scale, and (3) had the smallest standard deviation. To obtain androgynous voices, I selected vocal stimuli with ratings that (1) had the smallest effect size and, when possible, (2) did not significantly differ from the mid-point of the scale. Based on these criteria, I selected F1, F2, F3, F5; M1, M2, M3, M5; and A1, A2, A3, A5 as stimuli for use in future studies.

Table 2

Gender categorization, prototypicality, and gender typicality ratings for feminine voices

Note. Selected voices are bolded.

Voice	Pitch frequency	N	Gender categorization			<i>p</i> -value
			Percent man	Percent woman	Percent “unsure”	
F1	216.89	130	1.50%	98.50%	0%	<.001
F2	200.44	128	0.80%	99.20%	0%	<.001
F3	211.91	128	0%	99.20%	0.80%	<.001
F4	235.07	132	0%	99.20%	0.80%	<.001
F5	209.50	128	0%	99.20%	0.80%	<.001
Voice	Pitch frequency	N	Prototypicality ratings			<i>p</i> -value
			<i>M</i>	<i>SD</i>	Cohen's <i>d</i>	
F1	216.89	130	6.07	0.98	2.11	<.001
F2	200.44	128	6.12	0.99	2.15	<.001
F3	211.91	128	6.20	0.95	2.31	<.001
F4	235.07	132	5.92	1.14	1.69	<.001
F5	209.50	128	6.36	0.74	3.18	<.001
Voice	Pitch frequency	N	Gender typicality ratings			<i>p</i> -value
			<i>M</i>	<i>SD</i>	Cohen's <i>d</i>	
F1	216.89	130	6.01	0.94	2.13	<.001
F2	200.44	128	6.07	0.94	2.20	<.001
F3	211.91	128	6.05	0.92	2.22	<.001
F4	235.07	132	6.12	1.06	2.01	<.001
F5	209.50	128	6.23	0.78	2.87	<.001

Table 3*Gender categorization, prototypicality, and gender typicality ratings for masculine voices*

Note. Selected voices are bolded.

Voice	Pitch frequency	N	Gender categorization			p-value
			Percent man	Percent woman	Percent “unsure”	
M1	113.3	128	98.40%	1.60%	0%	<.001
M2	115.9	131	100.00%	0%	0%	<.001
M3	116.02	127	100.00%	0%	0%	<.001
M4	148.72	133	99.20%	0.80%	0%	<.001
M5	116.31	127	98.40%	1.60%	0%	<.001
Voice	Pitch frequency	N	Prototypicality ratings			p-value
			<i>M</i>	<i>SD</i>	Cohen's <i>d</i>	
M1	113.3	128	6.00	1.14	1.76	<.001
M2	115.9	131	6.17	1.04	2.09	<.001
M3	116.02	127	5.60	1.39	1.15	<.001
M4	148.72	133	5.75	1.27	1.38	<.001
M5	116.31	127	5.74	1.27	1.37	<.001
Voice	Pitch frequency	N	Gender typicality ratings			p-value
			<i>M</i>	<i>SD</i>	Cohen's <i>d</i>	
M1	113.3	128	2.48	1.41	-1.07	<.001
M2	115.9	131	2.25	1.43	-1.22	<.001
M3	116.02	127	2.79	1.51	-0.80	<.001
M4	148.72	133	2.44	1.33	-1.17	<.001
M5	116.31	127	2.61	1.36	-1.02	<.001

Table 4

Gender categorization, prototypicality, and gender typicality ratings for androgynous voices

Note. Selected voices are bolded.

Voice	Pitch frequency	N	Gender categorization			<i>p</i> -value
			Percent man	Percent woman	Percent “unsure”	
A1	176.55	131	23.70%	60.30%	16.00%	<.001
A2	167.24	132	43.20%	40.90%	15.90%	<.001
A3	162.87	127	48.80%	33.90%	17.30%	<.001
A4	168.53	125	52.80%	34.40%	12.80%	<.001
A5	166.34	131	57.30%	29.80%	13.00%	<.001
Voice	Pitch frequency	N	Prototypicality ratings			<i>p</i> -value
			<i>M</i>	<i>SD</i>	Cohen's <i>d</i>	
A1	176.55	131	3.82	1.45	-0.13	.190
A2	167.24	132	3.77	1.52	-0.15	.121
A3	162.87	127	4.26	1.57	0.16	.096
A4	168.53	125	4.07	1.53	0.05	.618
A5	166.34	131	3.97	1.60	-0.02	.861
Voice	Pitch frequency	N	Gender typicality ratings			<i>p</i> -value
			<i>M</i>	<i>SD</i>	Cohen's <i>d</i>	
A1	176.55	131	4.21	1.18	0.18	.040
A2	167.24	132	4.17	1.30	0.13	.142
A3	162.87	127	4.09	1.15	0.07	.400
A4	168.53	125	3.68	1.19	-0.27	.003
A5	166.34	131	4.12	1.40	0.09	.319

Question rating task

Domain specific questions. To select domain specific questions for use in future studies, I used a multi-step selection process (see Tables 5-6 for results). First, I used Chi-square tests to test whether the observed sample distribution significantly differed from the expected distribution based on chance for each question (i.e., 33.33%). Specifically, I examined whether the observed proportions of people who indicated “a man is more likely”, “a woman is more likely”, or “men and woman are equally likely” to answer the question significantly differed from 33.33%. Questions were selected if they were perceived as significantly more likely to be answered by a man or, separately, by a woman. Through this process, I selected questions 1-14 in the feminine domain, and questions 1-15 in the masculine domain.

Second, I examined whether gender ratings of feminine questions 1-14 and masculine questions 1-15 correlated with participant sex (effect coded as female = -1, male = 1) and political ideology. Given that participant sex and/or political ideology are potential variables of interest in examining preference for gender stereotypicality in future studies, I also sought to select questions with the smallest correlations between gender ratings and participant sex and political ideology. Finally, I only selected feminine and masculine questions that matched on their perceived difficulty based on visual inspection. Based on these criteria, I selected the following for masculine questions: “What is a hook angle in wood working?” ($M_{\text{difficulty}} = 4.27$, $M_{\text{interest}} = 3.12$, $M_{\text{curiosity}} = 3.07$), “What title is given to the NBA player if he receives the Maurice Podoloff Trophy?” ($M_{\text{difficulty}} = 4.08$, $M_{\text{interest}} = 3.05$, $M_{\text{curiosity}} = 2.96$), and “Which baseball team won the World Series in 1992?” ($M_{\text{difficulty}} = 4.14$, $M_{\text{interest}} = 2.77$, $M_{\text{curiosity}} = 2.65$). I selected the following for feminine questions: “What is the ballet term for spinning on one foot?” ($M_{\text{difficulty}} = 4.03$, $M_{\text{interest}} = 3.61$, $M_{\text{curiosity}} = 3.57$), “Which designer is famous for the wrap dress silhouette?”

($M_{\text{difficulty}} = 4.29$, $M_{\text{interest}} = 2.92$, $M_{\text{curiosity}} = 2.65$), and “Who is the Editor-in-Chief of the fashion magazine “Elle”?” ($M_{\text{difficulty}} = 4.09$, $M_{\text{interest}} = 2.71$, $M_{\text{curiosity}} = 2.54$).

Table 5

Gender ratings and correlations with participant sex and political ideology for feminine questions (Pilot Study 2)

Note. Selected questions are bolded. Questions are ordered by column “percent woman more likely” (from highest to lowest). Only correlations between gender ratings of questions 1-14, participant sex and political ideology are shown, as those questions were preselected based on their gender ratings. “#” indicates question number. * $p < .05$

#	Question	N	Gender Ratings				Correlations with Gender Ratings	
			% man more likely	% woman more likely	% man and woman equally likely	p -value	Participant Sex	Political Ideology
1	What color concealer works best to hide pimples and redness in the face?	109	0.00%	95.41%	4.59%	<.001	.12	.21*
2	What is the nail polish called that is colorless and goes on top of a colored nail polish?	105	2.86%	89.52%	7.62%	<.001	-.04	.10
3	Which designer is famous for the wrap dress silhouette?	107	2.80%	85.98%	11.21%	<.001	.06	.04
4	How many foot positions are there in classical Ballet?	107	0.93%	84.11%	14.95%	<.001	.23*	.11

Table 5 (cont.)

5	What type of knitting does the term "English knitting" describe?	112	1.79%	83.04%	15.18%	<.001	-.23	.001
6	What is the ballet term for spinning on one foot?	108	3.70%	82.41%	13.89%	<.001	-.02	-.06
7	What is the average age for babies to start crawling?	109	0.92%	81.65%	17.43%	<.001	-.13	-.11
8	Which famous supermodel tripped at her first fashion show?	109	0.92%	81.65%	17.43%	<.001	.03	.12
9	Before giving heated milk to a baby, it should always be tested against what?	108	3.70%	78.70%	17.59%	<.001	.03	.11
10	Who is the Editor-in-Chief of the fashion magazine "Elle"?	109	2.75%	77.98%	19.27%	<.001	.07	-.13
11	What cooking product can be added to the washing machine to reduce lint?	108	4.63%	76.85%	18.52%	<.001	-.03	.24*
12	How many pores are on the average adult female face?	107	0.93%	68.22%	30.84%	<.001	.17	-.01
13	Which automatic reflex does not occur in infants?	109	0.92%	61.47%	37.61%	<.001	-.02	-.08

Table 5 (cont.)

14	What is the best condiment to use when cleaning silver?	108	6.48%	59.26%	34.26%	<.001	-.10	.03
15	What is the process of washing wool called?	111	4.50%	53.15%	42.34%	<.001	-	-
16	Which artist composed the two ballets Sleeping Beauty and The Nutcracker?	110	0.91%	50.00%	49.09%	<.001	-	-
17	What development stage do children enter when they are in middle school?	107	0.93%	49.53%	49.53%	<.001	-	-
18	What is the best brush to use when cleaning the base of the faucet?	107	17.76%	38.32%	43.93%	<.001	-	-
19	What does "flipping a classroom" entail as an instructional strategy?	108	1.85%	30.56%	67.59%	<.001	-	-
20	White potatoes are high in what vitamin?	105	2.86%	27.62%	69.52%	<.001	-	-
21	Which nut has the most omega-3 fatty acids?	108	4.63%	25.00%	70.37%	<.001	-	-

Table 5 (cont.)

22	What is the daily maximum amount of sodium (in milligrams) recommended for a healthy adult?	107	1.87%	19.63%	78.50%	<.001	-	-
23	Which category of library materials do "almanacs" fall under?	108	5.56%	19.44%	75.00%	<.001	-	-
24	Which branch of psychology is devoted to the study of changes in behavior and abilities over the course of a lifetime?	110	4.55%	19.09%	76.36%	<.001	-	-
25	What is the printing method called, in which prints are made by pressing paper with an ink stone?	109	19.27%	15.60%	65.14%	<.001	-	-
26	What year was the Dewey Decimal System started?	108	7.41%	14.81%	77.78%	<.001	-	-
27	Which is the largest library in the world?	104	4.81%	14.42%	80.77%	<.001	-	-
28	The term "cultural literacy" coined by an American educator describes what?	105	5.71%	14.29%	80.00%	<.001	-	-

Table 5 (cont.)

29	The psychological test of human emotions and personality, using inkblots, is formally known as what?	107	6.54%	7.48%	85.98%	<.001	-	-
30	Which psychologist investigated obedience using electric shocks?	105	12.38%	4.76%	82.86%	<.001	-	-

Table 6

Gender ratings and correlations with participant sex and political ideology for masculine questions (Pilot Study 2)

Note. Selected questions are bolded. Questions are ordered by column “percent man more likely” (from highest to lowest). Only correlations between gender ratings of questions 1-15, participant sex and political ideology are shown, because those questions were preselected based on their gender ratings. “#” indexes question number. * $p < .05$; ** $p < .01$

#	Question	N	Gender Ratings				Correlations with Gender Ratings	
			% man more likely	% woman more likely	% man and woman equally likely	p -value	Participant Sex	Political Ideology
1	What is a hook angle in wood working?	106	88.68%	2.83%	8.49%	<.001	.12	-.01
2	What title is given to the NBA player if he receives the Maurice Podoloff Trophy?	108	87.96%	2.78%	9.26%	<.001	-.04	.09
3	How many engines did the Howard Hughes H-4 Hercules (The Spruce Goose) have?	108	85.19%	0.93%	13.89%	<.001	.04	.21*
4	What are the measurements of a standard size sheet of plywood, in inches?	109	84.40%	3.67%	11.93%	<.001	-.03	.08

Table 6 (cont.)

5	How many football teams are currently in the NFL?	108	84.26%	1.85%	13.89%	<.001	.17	-.02
6	Which type of fishing reel is most frequently used for beginners?	108	83.33%	2.78%	13.89%	<.001	.12	-.01
7	Which baseball team won the world series in 1992?	105	81.90%	2.86%	15.24%	<.001	.04	.05
8	What was the first car to break the sound barrier?	110	78.18%	2.73%	19.09%	<.001	.02	.03
9	What is a c-clamp in wood working used for?	108	77.78%	3.70%	18.52%	<.001	.32**	.05
10	What does the DOT number on a car tire stand for?	107	77.57%	2.80%	19.63%	<.001	.22*	.06
11	What are "fishing waders"?	105	70.48%	1.90%	27.62%	<.001	.16	.11
12	Dimples on golf balls help increase speed by reducing what?	110	68.18%	4.55%	27.27%	<.001	-.06	.16
13	Which component in a typical staircase is responsible for supporting the surface portion of stairs on which people walk?	112	66.96%	0.89%	32.14%	<.001	.03	.01

Table 6 (cont.)

14	Who designed the world's first car?	108	62.96%	3.70%	33.33%	<.001	-.08	.17
15	What are the four forces needed to make an airplane fly?	107	57.94%	2.80%	39.25%	<.001	-.05	.12
16	The massive four engine, double decker Airbus A380 is the world's largest passenger plane. How many passenger seats does it have?	106	50.00%	3.77%	46.23%	<.001	-	-
17	How many foxes, hares, or game birds are in a "brace"?	107	47.66%	6.54%	45.79%	<.001	-	-
18	What do deer shed every year?	106	47.17%	1.89%	50.94%	<.001	-	-
19	Which American architect has been called the "father of skyscrapers" and "father of modernism"?	109	43.12%	1.83%	55.05%	<.001	-	-
20	What is the smallest deer species in North America?	105	42.86%	4.76%	52.38%	<.001	-	-

Table 6 (cont.)

21	Which architecture term describes the interior curve of an arch?	108	39.81%	4.63%	55.56%	<.001	-	-
22	What is the largest species of fish in the world?	106	31.13%	5.66%	63.21%	<.001	-	-
23	What mathematical symbol was determined by Whiz Ferdinand von Lindemann in 1882?	107	30.84%	6.54%	62.62%	<.001	-	-
24	What was the name of the first artificial satellite launched into space?	103	30.10%	2.91%	66.99%	<.001	-	-
25	What is Newton's second law of motion?	109	23.85%	2.75%	73.39%	<.001	-	-
26	What type of galaxy is the most common in the universe?	107	22.43%	6.54%	71.03%	<.001	-	-
27	What is a body's resistance to changes in motion or speed called?	109	21.10%	4.59%	74.31%	<.001	-	-
28	What is the most common type of star found in the Milky Way?	106	20.75%	4.72%	74.53%	<.001	-	-

Table 6 (cont.)

29	What is an angle more than 90 degrees and less than 180 degrees called?	108	14.81%	6.48%	78.70%	<.001	-	-
30	What is the next prime number after 151?	110	14.55%	8.18%	77.27%	<.001	-	-

Pilot Study 2 Summary

In the second pilot study, I selected three masculine and three feminine questions, as well as four masculine, four feminine, and four androgynous voices for use in the main studies.

CHAPTER 3: PREFERENCE FOR GENDER STEREOTYPICALITY (STUDIES 1 & 2)

Study 1

In Study 1, I examined whether people prefer gender stereotypicality in voice-based AI (*gender stereotypicality hypothesis*). I predicted that participants would prefer gender stereotypicality (versus counterstereotypicality) in voice-based AI. Specifically, I predicted that participants would show an increased preference for a feminine-voiced (over masculine-voiced) AI when asked to answer a question in a stereotypically feminine (versus masculine) domain. The pre-registration plan for the hypotheses, methods, and analyses can be found here: https://osf.io/v8q43?view_only=6ce5d00e5c17419f8f08de8aba70c1f4

Method

Participants

351 participants (181 female participants, 165 male participants, 3 “Other” sex, 2 “Prefer not to answer”; $M_{\text{age}} = 40.26$; range = 19-77) were recruited from the MTurk website using the CloudResearch Approved Participant List. Seven additional participants completed the study, but were excluded from analyses for failing an attention check ($N = 5$), failing a soundcheck ($N = 1$), or not agreeing to have their data included ($N = 1$). According to a power analysis for crossed random effects (Westfall et al., 2014), a sample of at least 326 participants would provide at least 80% power to detect an effect size as small as $r = .15$. I anticipated that 10% of participants would fail an attention check or sound check, and so would be excluded from analysis. Thus, I collected 360 participants. All measures, manipulations, and exclusions in the study are disclosed.

Procedure

After providing consent, participants completed the same sound check as in Pilot Study 2. Participants then performed the voice preference task described below. Participants also filled out information about their prior AI use (see below). Then, participants provided basic demographic information, such as their age, race, and gender. Participants completed an attention check and were debriefed. Because the study involved deception, participants were asked to consent at the end of the study to have their data included in analyses.

Measures

Voice preference task. To examine preference for gender stereotypicality in AI, participants completed a voice preference task. In this task, participants first read the following instructions:

“On the following pages, you will listen to newly developed voice-based artificial intelligence (AI). We are interested in your preferences! AI involves computational machines that are intelligent and act like humans. Voice-based AI can process language and make decisions that normally require a human level of expertise. Sometimes AI is also used to help people answer questions. On the following pages, you will listen to different pairs of voice-based AI that have been randomly selected from a broad pool of AI. In each trial, you will be presented with two buttons that allow you to listen to short recordings of different voice-based AI. Each AI will say one example sentence. You can replay recordings as many times as needed by clicking on the buttons. Then, you will see a question. We would like for you to imagine that you could use one of the two voice-based AI to help you find an answer to the question. You will be asked to indicate which one of the voice-based AI you personally would like to choose to help you find an answer to the question. In total, you will complete this process four times.”

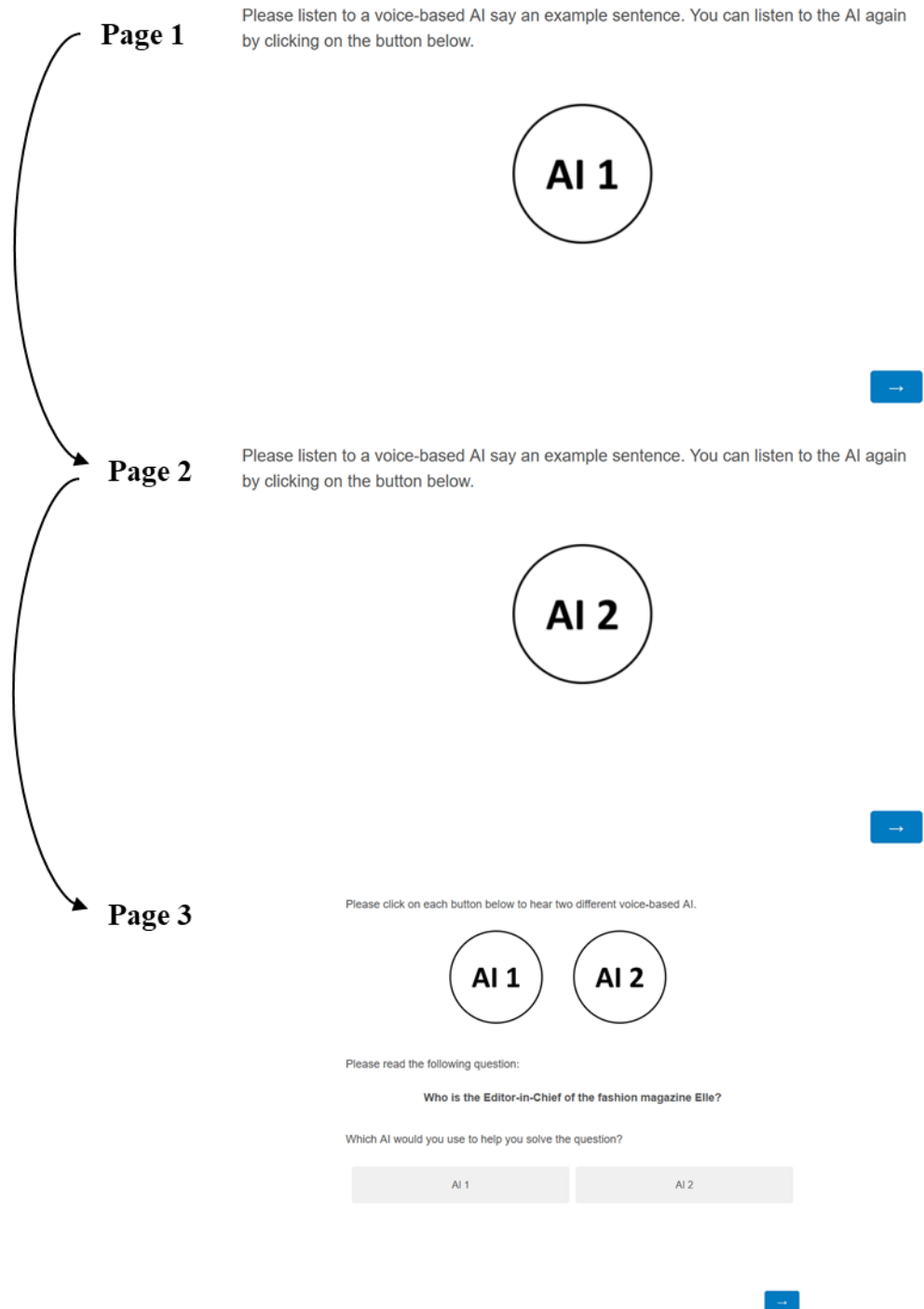
Next, participants completed four trials. Each trial was presented in one Qualtrics block. An example Qualtrics block is shown in Figure 2. There were a total of four feminine-voiced AI and four masculine-voiced AI. In each trial, participants were randomly presented with one feminine-voiced and one masculine-voiced AI. Each AI said one of three sentences selected at random. Participants were shown a button (i.e., “AI 1” or “AI 2”) and a vocal recording automatically played. Participants were able to replay the recording by clicking on the button. The order in which feminine- and masculine-voiced AIs were presented was counter-balanced, and voices were only shown once. The order of sentences spoken by each AI were randomly selected and voices could be presented more than once.

Participants were then shown the two AI buttons side-by-side and could replay the vocal recordings. They were also presented with one domain-specific question (e.g., Which baseball team won the world series in 1992?). Throughout the course of the four trials, participants were randomly assigned to see two of the three possible masculine questions and two of the three possible feminine questions.

Next, participants completed the main task of the experiment. Participants were asked “Which AI would you use to help solve the question” and made a binary choice between the masculine-voiced or feminine-voiced AI. Participants relied only on the AI’s vocal cues as a basis for their judgments. Participants were given as much time as needed to make their judgment. Participants repeated this process for the remaining three trials.

Figure 2

Example Qualtrics Block illustrating the Voice Preference Task



Type of prior AI usage. After completing the four trials, participants indicated their prior usage of AI in response to the question “What voice-based artificial intelligence do you use in everyday life?”. The options were: Amazon’s Alexa, Microsoft’s Cortana, Apple’s Siri, Google’s Assistant, I do not use voice-based artificial intelligence, Other (open-ended). This item was used for exploratory purposes. Participants who indicated that they do not use voice-based AI were not shown the subsequent frequency and prior use questions.

Frequency of prior AI usage. Participants responded to the question “How frequently do you use voice-based artificial intelligence?” on a 7-point scale (1 = *Not at all*, 7 = *All the time*). This item was used for exploratory analyses.

Prior AI use (open-ended). Participants’ modus of AI use was assessed using a single item created by the researcher. Participants were asked “What do you use voice-based artificial intelligence most frequently for?” and given space to provide a free written response. This item was used for exploratory purposes.

Results

Analytic strategy

In all analyses, alpha was set to .05. To test whether people preferred gender stereotypicality over counterstereotypicality in AI, I conducted generalized linear mixed-effects models. For all models, I used the `glmer` function from the `lme4` package (Bates et al., 2015) in R to account for nonindependence in judgments (Fitzmaurice et al., 2012). In all models, I specified random intercepts of participant, feminine AI voice, masculine AI voice, question, gender-neutral sentence said by a feminine voice, and gender-neutral sentence said by a masculine voice to account for between-participant and between-stimuli variance (Judd et al.,

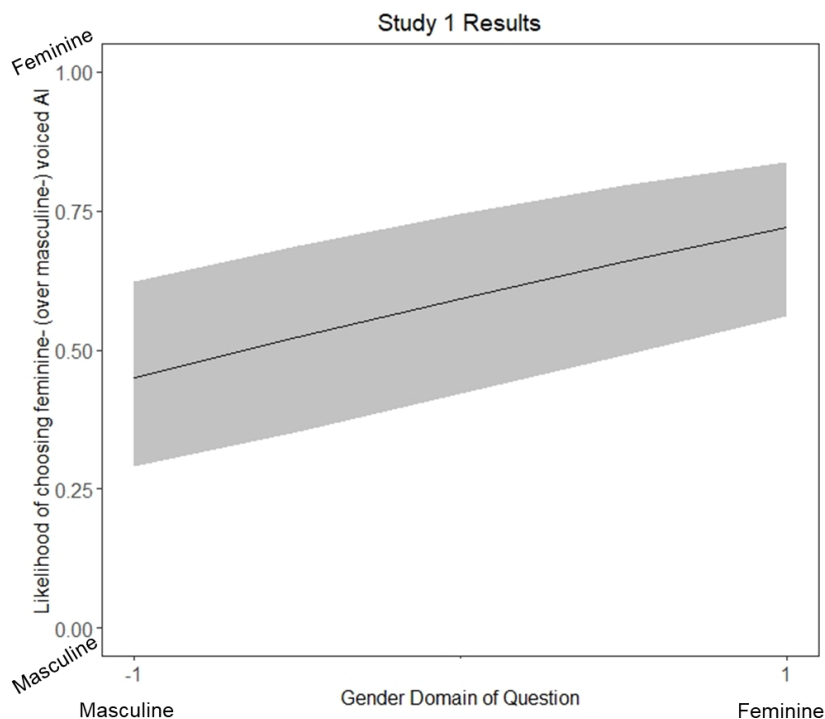
2012). Preferred AI was the dependent variable in all models and specified as a binary logistic outcome (coded as masculine = 0, feminine = 1).

Do people prefer gender stereotypicality in AI?

I specified a model that included the fixed effect of gender domain of question (effect coded as feminine = 1, masculine = -1) as the predictor variable. There was a main effect of gender domain of question, $B = 0.57$, $SE = 0.06$, Wald $\chi^2 = 94.13$, $p < .001$, 95% CI [0.45, 0.68], odds ratio = 1.77, indicating that the likelihood of choosing a feminine- (over masculine-) voiced AI was higher when asked to answer a question in a stereotypically feminine (versus masculine) domain (see Figure 3). This finding supports the prediction that people will prefer gender stereotypicality in AI.

Figure 3

Likelihood of Choosing Feminine- (Over Masculine-) Voiced AI as a Function of Gender Domain of Question (Study 1).



Does the effect of gender stereotypicality emerge in both masculine and feminine domains?

I next conducted exploratory analyses examining whether the preference for gender stereotypicality in AI occurred in both feminine and masculine domains.

Feminine domain. To test the stereotypicality effect within the feminine domain, I specified the same model outlined above. The fixed effect predictor of gender domain was dummy coded as feminine = 0 and masculine = 1. The intercept was significant and positive, $B = 0.95$, $SE = 0.36$, Wald $\chi^2 = 7.05$, $p = .008$, 95% CI [0.25, 1.65], odds ratio = 2.58, indicating that the likelihood of choosing a feminine- (versus masculine-) voiced AI was higher when presented with feminine domain questions.

Masculine domain. To test the stereotypicality effect within the masculine domain, I specified the same model outlined above. The fixed effect predictor of gender domain was dummy coded as masculine = 0 and feminine = 1. The intercept was not significant, $B = -0.19$, $SE = 0.36$, Wald $\chi^2 = 0.29$, $p = .592$, 95% CI [-0.89, 0.51], odds ratio = 0.83, indicating that the likelihood of choosing a masculine or feminine-voiced AI did not significantly differ when presented with masculine domain questions.

Participant sex as a moderator

I next conducted an exploratory analysis examining whether participant sex impacted the strength of the stereotypicality preference effect. For the purpose of this analysis, I excluded participants who identified their sex as non-binary (“Other”; $N = 3$) or did not report their sex ($N = 2$). The fixed effects were the gender domain of question (effect coded as masculine = -1, feminine = 1), participant sex (effect coded as female = -1, male = 1), and their interaction. The fixed effect of gender domain was significant, $B = 0.56$, $SE = 0.06$, Wald $\chi^2 = 88.13$, $p < .001$, 95% CI [0.44, 0.67], odds ratio = 1.74, indicating that the likelihood of choosing a feminine-

(over masculine-) voiced AI was higher when asked to answer a question in a stereotypically feminine (versus masculine) domain. The fixed effect of participant sex was also significant, $B = 0.15$, $SE = 0.06$, Wald $\chi^2 = 6.28$, $p = .012$, 95% CI [0.03, 0.28], odds ratio = 1.17, indicating that the likelihood of choosing a feminine-voiced (over masculine-) voiced AI was higher for female than for male participants. The interaction was also significant, $B = -0.13$, $SE = 0.06$, Wald $\chi^2 = 5.01$, $p = .025$, 95% CI [-0.25, -0.02], odds ratio = 0.88, indicating that the likelihood of choosing a feminine- (over masculine-) voiced AI based on gender domain differed across participant sex. Importantly, however, the effect of gender domain of question was significant both among female participants, $B = 0.69$, $SE = 0.08$, Wald $\chi^2 = 70.39$, $p < .001$, 95% CI [0.53, 0.85], odds ratio = 1.99, and male participants, $B = 0.42$, $SE = 0.09$, Wald $\chi^2 = 24.93$, $p < .001$, 95% CI [0.26, 0.59], odds ratio = 1.53. These findings indicate that both female and male participants preferred gender stereotypicality in AI, but that this preference was stronger among female participants.

Frequency of prior AI usage as a moderator

I next conducted an exploratory analysis examining whether frequency of prior AI usage impacted the strength of the stereotypicality effect. The fixed effects were the gender domain of the question (effect coded as masculine = -1, feminine = 1), AI usage (grand mean centered), and their interaction. The fixed effect of gender domain was significant, $B = 0.60$, $SE = 0.07$, Wald $\chi^2 = 80.98$, $p < .001$, 95% CI [0.47, 0.73], odds ratio = 1.82, indicating that the likelihood of choosing a feminine- (over masculine-) voiced AI was higher when asked to answer a question in a stereotypically feminine (versus masculine) domain. Neither the main effect of AI usage, $B = 0.03$, $SE = 0.04$, Wald $\chi^2 = 0.42$, $p = .518$, 95% CI [-0.05, 0.10], odds ratio = 1.03, nor the interaction between gender domain and AI usage were significant, $B = -0.03$, $SE = 0.04$, Wald χ^2

= 0.56, $p = .457$, 95% CI [-0.11, 0.05], odds ratio = 0.97. These findings indicate that the preference for gender stereotypicality in AI did not vary based on frequency of prior AI usage.

Study 1 Summary

Participants showed an increased preference for a feminine-voiced (over masculine-voiced) AI when asked to answer a question in a stereotypically feminine (versus masculine) domain. Exploratory analyses indicated that this preference for gender stereotypicality (versus counterstereotypicality) emerged in the feminine domain but not in the masculine domain. Further, the effect was stronger among female than male participants, but did not vary across people with more or less frequent AI usage.

Study 2

In Study 2, I sought to replicate and extend the findings from Study 1. Specifically, I examined how the option of an androgynous-voiced AI (i.e., blend of prototypically feminine- and masculine-voices) would affect people's preferences. Androgynous individuals possess both feminine and masculine characteristics (Bem, 1981a). In other words, they occupy a space in which they are never entirely stereotypical nor counterstereotypical. Extant literature outlines competing ideas regarding whether people might prefer gender stereotypicality or androgyny in AI. To the extent that gender stereotypes are universally used, I predict that people will prefer gender stereotypicality (compared to counterstereotypicality and androgyny) in voice-based AI (e.g., increased preference for a feminine- (compared to masculine- or androgynous-) voiced AI in a stereotypically feminine (versus masculine) domain). This pattern would lend support for the *gender stereotypicality hypothesis* and to the idea that the gender binary is rigidly applied to AI.

At the same time, research based on psychological and physical androgyny in humans outlines ideas that would support a competing prediction to the gender stereotypicality

hypothesis. Some work finds that psychologically and physical androgynous individuals are perceived more favorably than gender stereotypical individuals (Basow & Howe, 1987; Hester et al., 2021; Jackson, 1983; Major et al., 1981). Non-human entities without biological sex (e.g., AI) might prompt a preference for voices that diverge from prototypical gender expectations (i.e., androgyny). Thus, people may prefer an androgynous- (over feminine- or masculine-) voiced AI regardless of gender domain. Specifically, participants may show an increased preference for an androgynous- (compared to feminine- and masculine-) voiced AI regardless of whether the domain is stereotypically masculine or feminine (*androgyny preference prediction*).

Although some evidence suggests that people might prefer androgyny (over stereotypicality) in AI, work on perceptual fluency also suggests that people might most be most inclined to dislike androgyny. The perceived ease with which people process available information guides evaluative judgments (Winkielman et al., 2006). Stimuli that are processed more easily and faster are rated as more favorable than disfluently processed stimuli (Alter & Oppenheimer, 2006; Belke et al., 2010; Winkielman et al., 2006). Ambiguity in social categories that push beyond single identity frameworks in race, gender, or sexuality is perceived more negatively (Lick et al., 2015; Remedios et al., 2012; Sanchez & Bonam, 2009). In addition, people are evaluated more negatively when their gendered characteristics are more androgynous because they are more difficult to place into a social category (Atwood & Axt, 2021; Lick & Johnson, 2013, 2015; Stern & Rule, 2018). According to this line of reasoning, androgyny might be preferred less than gender stereotypicality even in the context of AI. This prediction would complement the gender stereotypicality hypothesis and suggests that people would show the lowest preference for androgyny in voice-based AI (*androgyny derogation prediction*). Specifically, people might show a decreased preference for androgyny compared to both

stereotypicality and counterstereotypicality in voice-based AI regardless of whether the domain is stereotypically masculine or feminine.

Taken together, in Study 2, I advance the literature through examining people's preference for stereotypicality (compared to counterstereotypicality and androgyny) in AI. The pre-registration plan for the hypotheses, methods, and analyses can be found here:

https://osf.io/zfwne/?view_only=c2ee67a6a8c14c84808d50e6aa5d6be8

Method

Participants

488 participants (263 female participants, 218 male participants, 2 “Other” sex, 5 “Prefer not to answer”; $M_{age} = 40.40$; range = 19-80) were recruited from Mturk using the CloudResearch Approved Participant List. Ten additional participants completed the study, but were excluded from analyses for failing an attention check ($N = 8$) or failing a soundcheck ($N = 2$). According to a power analysis for crossed random effects (Westfall et al., 2014), a sample of at least 326 participants would provide at least 80% power to detect an effect size as small as $r = .15$. I expected approximately 10% of participants to fail an attention check. To amplify power and account for attention checks, I planned to collect more participants than deemed needed by the power analysis. Thus, I aimed to collect 500 participants.

Procedure

The procedure was the same as in Study 1. The voice preference task was amended to add a third type of voice (i.e., androgynous).

Voice preference task. To examine preference for gender stereotypicality in AI, participants completed the same voice preference task as in Study 1, except participants were

shown three AI instead of two. Specifically, participants were shown a feminine-, masculine-, and androgynous-voiced AI in each of the four trials.

Results

Analytic strategy

In all analyses, alpha was set to .05. For the purposes of these analyses, I created a new variable indicating preference for stereotypicality (e.g., choosing feminine-voiced AI in feminine domain; coded as 0), counterstereotypicality (e.g., choosing masculine-voiced AI in a feminine domain; coded as 1), or androgyny (e.g., choosing androgynous-voiced AI in a feminine domain; coded as 2). For all models, I used the glmer function from the lme4 package (Bates et al., 2015) in R to account for nonindependence in judgments (Fitzmaurice et al., 2012). In all models, I specified random intercepts of participant, feminine AI voice, masculine AI voice, androgynous AI voice², question, gender-neutral sentence said by an androgynous voice², gender-neutral sentence said by a feminine voice, and gender-neutral sentence said by a masculine voice to account for between-participant and between-stimuli variance (Judd et al., 2012). I conducted separate binary logistic models when comparing stereotypicality to counterstereotypicality and stereotypicality to androgyny (it is not possible at the current time to run a multilevel multinomial model with several random effects in R using CRAN).

Do people prefer gender stereotypicality in AI?

I first conducted confirmatory analyses examining whether the gender stereotypicality hypothesis was supported. For the purpose of this analysis, I conduct two intercept-only models. The first model compared preferences for stereotypicality against counterstereotypicality, and the dependent variable was coded as counterstereotypical = 0, stereotypical = 1. The intercept was

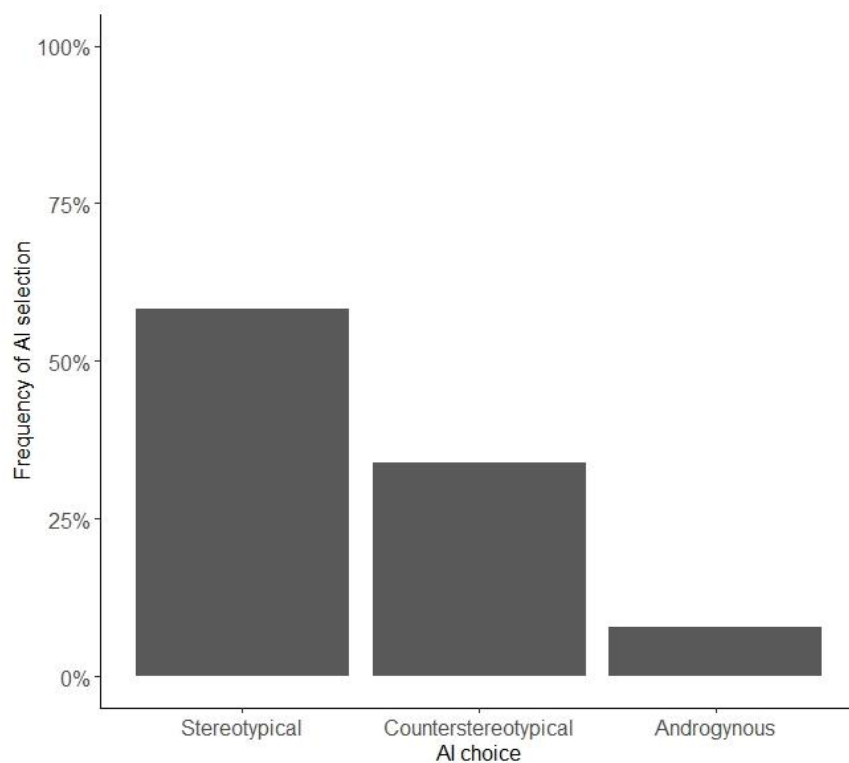
² There is an error in the preregistration, which does not specify a random intercept of androgynous AI voice and what sentence was said by the androgynous voice. I included those random effects here.

significant, $B = 0.59$, $SE = 0.22$, Wald $\chi^2 = 7.20$, $p = .007$, 95% CI [0.16, 1.02], odds ratio = 1.80, indicating that participants expressed a greater preference for the stereotypicality than counterstereotypicality.

The second model compared preferences for the stereotypicality against androgyny, and the dependent variable was coded as androgynous = 0, stereotypical = 1. The intercept was significant, $B = 2.18$, $SE = 0.38$, Wald $\chi^2 = 32.74$, $p < .001$, 95% CI [1.43, 2.93], odds ratio = 8.87, meaning that participants expressed a greater preference for the stereotypicality than androgyny. Overall, these findings support the gender stereotypicality hypothesis. Specifically, participants preferred gender stereotypicality more strongly than counterstereotypicality and androgyny in voiced-based AI (see Figure 4).

Figure 4

Frequencies of Choosing Stereotypical, Counterstereotypical, and Androgynous AI (Study 2).



Exploratory Analyses

In the next section, I report exploratory analyses examining whether people prefer androgyny least and whether the people's preferences vary as a function of the gender of the domain, participant sex, and frequency of prior AI usage.

Do people prefer androgyny in AI the least?

I next conducted exploratory analyses examining whether the androgyny derogation effect emerges. For the purpose of this analysis, I conduct two intercept-only models. The first model compared preferences for the androgyny against stereotypicality, and the dependent variable was coded as stereotypical = 0, androgynous = 1. The intercept was significant, $B = -2.18$, $SE = 0.38$, Wald $\chi^2 = 32.74$, $p < .001$, 95% CI [-2.93, -1.43], odds ratio = 0.11, indicating that (as noted above) participants expressed greater preference for stereotypicality than androgyny.

The second model compared preferences for androgyny against counterstereotypicality, and the dependent variable was coded as counterstereotypical = 0, androgynous = 1. The intercept was significant, $B = -1.61$, $SE = 0.41$, Wald $\chi^2 = 15.70$, $p < .001$, 95% CI [-2.40, -0.81], odds ratio = 0.20, indicating that participants expressed greater preference for counterstereotypicality than androgyny. Overall, these findings support the androgyny derogation prediction. Participants showed the lowest preference for androgyny in voice-based AI. Specifically, participants preferred androgyny the least compared to stereotypicality and counterstereotypicality in voice-based AI.

Does the effect of gender stereotypicality emerge in both feminine and masculine domains?

I next conducted exploratory analyses examining whether the preference for gender stereotypicality in AI varied across masculine and feminine domains. For the purpose of this

analysis, I conducted two models. The first model compared preferences for stereotypicality against counterstereotypicality. The fixed effect was the gender domain of the question (effect coded as feminine = 1, masculine = -1), and the dependent variable was coded as counterstereotypical = 0, stereotypical = 1. The effect of gender domain was significant, $B = 0.45$, $SE = 0.05$, Wald $\chi^2 = 78.22$, $p < .001$, 95% CI [0.35, 0.55], odds ratio = 1.57, indicating that the likelihood of preferring stereotypicality (versus counterstereotypicality) is higher in a feminine (versus masculine) domain.

The second model compared preferences for the stereotypicality against androgyny. The fixed effect was the gender domain of the question (effect coded as feminine = 1, masculine = -1), and the dependent variable was coded as androgynous = 0, stereotypical = 1. The effect of gender domain was significant, $B = 0.30$, $SE = 0.11$, Wald $\chi^2 = 7.02$, $p = .008$, 95% CI [0.08, 0.53], odds ratio = 1.36, indicating that the likelihood of preferring stereotypicality (versus androgyny) is also higher in a feminine (versus masculine) domain.

Given that effects of gender domain of question were significant when comparing stereotypicality versus counterstereotypicality and stereotypicality versus androgyny, I next tested the effect of gender stereotypicality preference within the feminine and masculine domain separately.

Feminine domain. For the feminine domain questions, the likelihood of preferring stereotypicality was higher than the likelihood of preferring counterstereotypicality, $B = 1.05$, $SE = 0.11$, Wald $\chi^2 = 96.29$, $p < .001$, 95% CI [0.84, 1.25], odds ratio = 2.85, and androgyny, $B = 2.48$, $SE = 0.37$, Wald $\chi^2 = 44.01$, $p < .001$, 95% CI [1.75, 3.21], odds ratio = 11.91. Overall, these findings support the gender stereotypicality hypothesis in the feminine domain.

Masculine domain. For the masculine domain questions, the likelihood of preferring stereotypicality was not different than the likelihood of preferring counterstereotypical AI, $B = 0.14$, $SE = 0.10$, Wald $\chi^2 = 1.89$, $p = .170$, 95% CI [-0.06, 0.34], odds ratio = 1.15, but was higher than the likelihood of preferring androgyny, $B = 1.87$, $SE = 0.37$, Wald $\chi^2 = 25.47$, $p < .001$, 95% CI [1.14, 2.59], odds ratio = 6.48. Overall, these findings partially support the gender stereotypicality hypothesis in the masculine domain.

Does the effect of androgyny derogation emerge in both feminine and masculine domains?

I next conducted exploratory analyses examining whether the androgyny derogation effect varied across masculine and feminine domains. For the purpose of this analyses, I conducted two models. The first model compared preferences for androgyny against stereotypicality. The fixed effect was the gender domain of the question (effect coded as feminine = 1, masculine = -1), and the dependent variable was coded as stereotypical = 0, androgynous = 1. The effect of gender domain was significant, $B = -0.30$, $SE = 0.11$, Wald $\chi^2 = 7.02$, $p = .008$, 95% CI [-0.53, -0.08], odds ratio = 0.74, indicating that the likelihood of preferring stereotypicality (over androgyny) is higher in a feminine than masculine domain.

The second model compared preferences for androgyny against counterstereotypicality. The fixed effect was the gender domain of the question (effect coded as feminine = 1, masculine = -1), and the dependent variable was coded as counterstereotypical = 0, androgynous = 1. The effect of gender domain was significant, $B = 0.24$, $SE = 0.12$, Wald $\chi^2 = 4.37$, $p = .037$, 95% CI [0.01, 0.47], odds ratio = 1.27, indicating that the likelihood of preferring counterstereotypicality (over androgyny) is higher in a feminine than masculine domain.

Given that both effects of gender domain were significant, I next tested the effect of androgyny derogation within the feminine and masculine domain separately.

Feminine domain. For the feminine domain questions, the likelihood of preferring androgyny was lower than the likelihood of preferring stereotypicality, $B = -2.48$, $SE = 0.37$, Wald $\chi^2 = 44.01$, $p < .001$, 95% CI [-3.21, -1.75], odds ratio = 0.08, or counterstereotypicality, $B = -1.34$, $SE = 0.41$, Wald $\chi^2 = 10.63$, $p = .001$, 95% CI [-2.15, -0.54], odds ratio = 0.26. Overall, these findings support the androgyny derogation prediction in the feminine domain.

Masculine domain. For the masculine domain questions, the likelihood of preferring androgyny was lower than the likelihood of preferring stereotypicality, $B = -1.87$, $SE = 0.37$, Wald $\chi^2 = 25.47$, $p < .001$, 95% CI [-2.59, -1.14], odds ratio = 0.15, or counterstereotypicality, $B = -1.82$, $SE = 0.41$, Wald $\chi^2 = 20.27$, $p < .001$, 95% CI [-2.62, -1.03], odds ratio = 0.16. Overall, these findings support the androgyny derogation prediction in the masculine domain.

Participant sex as a moderator

I next conducted exploratory analyses examining whether participant sex impacted the strength of the gender stereotypicality and androgyny derogation effects. For the purpose of these analyses, I excluded participants who identified their sex as non-binary (“Other”) or did not report their sex. All models included the fixed effect of participant sex (effect coded as female = -1, male = 1).

Participant sex as a moderator (gender stereotypicality hypothesis). The first model compared preferences for stereotypicality against those for counterstereotypicality, and the dependent variable was coded as counterstereotypical = 0, stereotypical = 1. The effect of participant sex was not significant, $B = -0.04$, $SE = 0.06$, Wald $\chi^2 = 0.50$, $p = .478$, 95% CI [-0.16, 0.08], odds ratio = 0.96, indicating that participant sex did not modulate how strongly people preferred stereotypicality over counterstereotypicality. The second model compared preferences for stereotypicality against those for androgyny, and the dependent variable was

coded as androgynous = 0, stereotypical = 1. The effect of participant sex was again not significant, $B = -0.05$, $SE = 0.10$, Wald $\chi^2 = 0.22$, $p = .638$, 95% CI [-0.24, 0.15], odds ratio = 0.95, indicating that participant sex did not modulate how strongly people preferred stereotypicality over androgyny. Overall, these findings indicate that participant sex did not impact the strength of the gender stereotypicality effect.

Participant sex as a moderator (androgyny derogation prediction). The first model compared preferences for androgyny against those for stereotypicality, and the dependent variable was coded as stereotypical = 0, androgynous = 1. The effect of participant sex was not significant, $B = 0.05$, $SE = 0.10$, Wald $\chi^2 = 0.22$, $p = .638$, 95% CI [-0.15, 0.24], odds ratio = 1.05, indicating that participant sex did not modulate how strongly people preferred androgyny versus stereotypicality. The second model compared preferences for androgyny against those for counterstereotypicality, and the dependent variable was coded as counterstereotypical = 0, androgynous = 1. The effect of participant sex was not significant, $B = -0.03$, $SE = 0.11$, Wald $\chi^2 = 0.10$, $p = .751$, 95% CI [-0.25, 0.18], odds ratio = 0.97, indicating that participant sex did not modulate how strongly people preferred androgyny over counterstereotypicality. Overall, these findings indicate that participant sex did not impact the strength of the androgyny derogation effect.

Frequency of prior AI usage as a moderator

I next conducted exploratory analyses examining whether frequency of prior AI usage impacted the strength of the gender stereotypicality and androgyny derogation effects. The fixed effect in all models was prior AI use (mean centered).

Frequency of prior AI usage as a moderator (gender stereotypicality hypothesis).

The first model compared preferences for stereotypicality against those for

counterstereotypicality, and the dependent variable was coded as counterstereotypical = 0, stereotypical = 1. The effect of prior AI use was not significant, $B = 0.02$, $SE = 0.04$, Wald $\chi^2 = 0.28$, $p = .595$, 95% CI [-0.06, 0.10], odds ratio = 1.02, indicating that frequency of AI use did not modulate how strongly people preferred stereotypicality over counterstereotypicality. The second model compared preferences for stereotypicality against those for androgyny, and the dependent variable was coded as androgynous = 0, stereotypical = 1. The effect of AI use frequency was not significant, $B = -0.02$, $SE = 0.07$, Wald $\chi^2 = 0.07$, $p = .793$, 95% CI [-0.16, 0.13], odds ratio = 0.98, indicating that frequency of AI use did not modulate how strongly people preferred stereotypicality over androgyny. Overall, findings indicated that frequency of prior AI usage did not impact the strength of the gender stereotypicality effect.

Frequency of prior AI usage as a moderator (androgyny derogation prediction). The first model compared preferences for androgyny against those for stereotypicality, and the dependent variable was coded as stereotypical = 0, androgynous = 1. The effect of prior AI use was not significant, $B = 0.02$, $SE = 0.07$, Wald $\chi^2 = 0.07$, $p = .793$, 95% CI [-0.13, 0.16], odds ratio = 1.03, indicating that frequency of AI use did not modulate how strongly people preferred androgyny over stereotypicality. The second model compared preferences for androgyny against those for counterstereotypicality, and the dependent variable was coded as counterstereotypical = 0, androgynous = 1. The effect of prior AI use was not significant, $B = 0.05$, $SE = 0.08$, Wald $\chi^2 = 0.49$, $p = .483$, 95% CI [-0.10, 0.20], odds ratio = 1.06, indicating that frequency of AI use did not modulate how strongly people preferred androgyny over counterstereotypicality. Overall, these findings indicate the frequency of prior AI use did not impact the strength of the androgyny derogation effect.

Study 2 Summary

I replicated and extended Study 1, finding that participants preferred gender stereotypicality (versus counterstereotypicality and androgyny) in voice-based AI. Similar to Study 1, the gender stereotypicality preference was stronger in the feminine domain compared to the masculine domain. Androgynous-voiced AI were preferred least. These effects did not vary between female and male participants, or people with more or less frequent AI usage.

CHAPTER 4: MECHANISM DRIVING PREFERENCE FOR GENDER STEREOTYPICALITY (STUDIES 3 & 4)

Study 3

In Study 3, I extended findings from Studies 1 and 2, and sought to examine perceived credibility as a potential mechanism explaining why people prefer stereotypicality in AI. Specifically, I investigated whether an AI is seen as more likely to provide the correct answer to a gender domain question that stereotypically (versus counterstereotypically) matches the gender of the AI voice (*credibility hypothesis*).

Technologies without biological sex, such as AI, do not (yet) have different knowledge or prior experiences based on gender. In line with this, an AI voice should objectively not impact perceived expertise. In other words, AI have equal knowledge regardless of the gender of the voice, thus would be able to answer any gendered question. However, people may perceive feminine- and masculine-voiced AI differentially capable of answering questions given their perceived stereotypical attributes. This reasoning is in line with social role theory, in which a central tenet posits that people form role expectations and desired abilities as they observe men and women engage in their respective social roles (Eagly & Steffen, 1984; Eagly, 1987). Thus, people believe that men and women are differentially capable of fulfilling a role or completing a domain specific task (Eagly & Karau, 2002). In other words, different roles or tasks are perceived as better suited for men or women given their stereotypical abilities. For instance, people hold stereotypical beliefs that women are better suited for a domestic role given its requirement for feminine characteristics, whereas men are seen as better suited for an employment role as it requires masculine characteristics (Eagly & Steffen, 1984). Further, domains such as science and math are stereotyped as masculine, whereas domains such as

teaching and humanities are stereotyped as feminine (Koenig et al., 2011; Nosek et al., 2002; Nosek et al., 2007; White & White, 2006). These stereotypical beliefs about men's and women's abilities also influence hiring decisions and occupational sex segregation (e.g., Cohen & Bunker, 1975; Lippa & Connelly, 1990). For example, incongruity between perceived abilities of a woman and a stereotypical masculine job results in perceived lack of fit and hiring discrimination (Glick et al., 1988).

To the extent that men and women stereotypically occupy different domains, greater gender role expectations (i.e., domain knowledge) are associated with men and women. In the *credibility hypothesis*, I predict that people will rate an AI voice as more credible when the gender of the AI voice is a stereotypical (versus counterstereotypical) match to the gender of the question (e.g., high perceived credibility of a feminine AI voice paired with a feminine question versus a masculine question). The pre-registration plan for the hypotheses, methods, and analyses can be found here:

https://osf.io/hdfqb/?view_only=f582eb83516d45ae92fb18e6f61d8d90

Method

Participants

479 participants (240 women, 227 men, 8 “Other”, 4 “Prefer not to answer”; $M_{\text{age}} = 41.50$; range = 18-78) were recruited from Mturk using the CloudResearch Approved Participant List. 19 additional participants completed the study, but were excluded from analyses for failing an attention check ($N = 15$), failing a soundcheck ($N = 3$), or not agreeing to have their data included ($N = 4$).³ According to a power analysis for crossed random effects (Westfall et al., 2014), a sample of at least 330 participants would provide at least 80% power to detect an effect

³ The number of excluded participants does not add to 100%, because some participants failed multiple exclusion criteria.

size as small as $r = .15$.⁴ I expected approximately 10% of participants to fail an attention check. To amplify power and account for attention checks, I planned to collect more participants than deemed needed by the power analysis. Thus, I aimed to collect 500 participants.

Procedure

After providing consent, participants completed the same sound check as in previous studies. Participants then performed the AI credibility task described below. After completing all six trials of the task, participants completed questions about frequency of AI use, a demographic questionnaire, an attention check, questions about the purpose of the study, and a debrief. Because the study involved deception, participants were asked to re-consent at the end of the study to have their data included in analyses.

AI credibility task. Participants were randomly assigned to complete six trials. Each trial was presented in a separate Qualtrics block. In each trial, participants were presented with one AI voice-question pairing. These pairings were either stereotypical (e.g., feminine-voiced AI paired with feminine domain question) or counterstereotypical (e.g., feminine-voiced AI paired with masculine domain question). This resulted in four possible AI voice and question pairings: (1) stereotypically feminine AI voice-question pairing (i.e., feminine voice, feminine question), (2) stereotypically masculine AI voice-question pairing (i.e., masculine voice, masculine question), (3) counterstereotypical feminine AI voice-question pairing (i.e., feminine voice, masculine question), and (4) counterstereotypical masculine AI voice-question pairing (i.e., masculine voice, feminine question).

Participants completed six trials total. Specifically, participants were presented with each of the above four pairings in random order. In addition, participants were presented with 2 more

⁴ There is a typo in the required sample in the preregistration. Instead of noting that I needed at least 330 participants, it states that I need at least 430 participants.

pairings. These two additional pairings were iterations of the four possible AI voice-question pairings listed above. These two pairings were evenly distributed across participants to ensure that all pairings were presented at similar rates. Additionally, other stipulations to the AI voice and question pairings were that (1) there was no repetition in the voices presented to participants, and (2) there was no repetition in the question presented to participants.

Each trial was presented on one Qualtrics block. An example Qualtrics block is shown in Figure 5. In each trial, participants were shown a button (e.g., “AI 1”), and a vocal recording would automatically play. Participants were randomly presented with a feminine-voiced or masculine-voiced AI saying a gender-neutral sentence. Participants were able to replay the recording by clicking on the button. On the same page, participants were randomly presented with one of three masculine or one of three feminine domain specific questions as stimuli (across the six trials, participants were presented with all six questions). Participants were then asked “How likely is the AI to correctly answer this question?” and responded on a 7-point scale (1 = *Not at all likely*, 4 = *Somewhat likely*, 7 = *Very likely*).

Demographics. Participants reported their age, race, and gender. Participants’ gender was assessed using the following question: “What is your gender?” (*Man, Woman, Other: [open ended], Prefer not to answer*).

Figure 5

Example Qualtrics Block illustrating the AI Credibility Task

Please listen to a voice-based AI say an example sentence. You can listen to the AI again by clicking on the button below.

AI 1

Please read the following question:

What is a hook angle in wood working?

How likely is the AI to correctly answer this question?

Not at all likely 1	2	3	Somewhat likely 4	5	6	Very likely 7
------------------------	---	---	----------------------	---	---	------------------

Results

Analytic strategy

In all analyses, alpha was set to .05. For all models, I used the lmer function from the lme4 package (Bates et al., 2015) in R to account for nonindependence in judgments (Fitzmaurice et al., 2012). In all models, I specified random intercepts of participant, voice, question, and gender-neutral sentence said by the voice to account for between-participant and between-stimuli variance (Judd et al., 2012). The dependent variable was perceived credibility in all models.

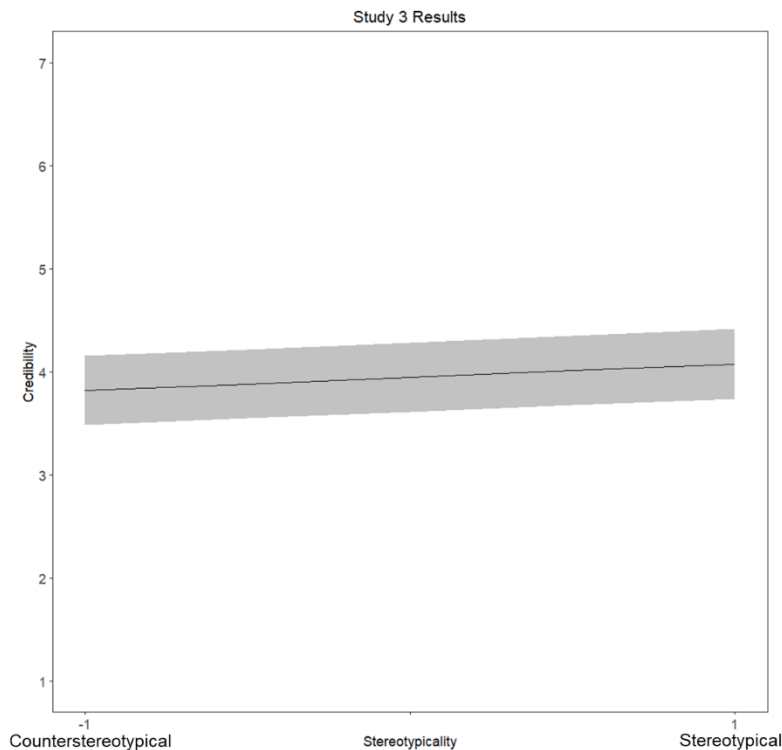
Is stereotypicality in AI perceived as more credible than counterstereotypicality?

I first conducted a confirmatory analysis examining whether the credibility effect emerges. I specified the fixed effect of stereotypicality (effect coded as counterstereotypical = -1, stereotypical = 1) as the predictor variable. The main effect of stereotypicality was significant, $B = 0.13$, $SE = 0.24$, $t(2415) = 5.39$, $p < .001$, 95% CI [0.09, 0.18], indicating that participants perceived an AI as more credible when the gender of the AI voice was a stereotypical (versus

counterstereotypical) match to the gender of the question (see Figure 6). This finding supports the credibility hypothesis.

Figure 6

Perceived Credibility as a Function of the Stereotypicality of AI (Study 3).



Exploratory Analyses

In the next section, I report exploratory analyses examining whether the credibility effect varies as a function of the gender of the AI voice, participant gender, and frequency of prior AI usage. To address potential demand characteristics of the study's design, I also assessed whether correctly guessing the hypothesis would modulate results.

Variation across gender of AI voice

I conducted an exploratory analysis examining whether the credibility effect varies across gender of AI voice. The fixed effects in the model were stereotypicality (effect coded as

counterstereotypical = -1, stereotypical = 1), gender of AI voice (effect coded as feminine = -1, masculine = 1), and their interaction. The fixed effect of stereotypicality was significant, $B = 0.13$, $SE = 0.24$, $t(2495) = 5.63$, $p < .001$, 95% CI [0.09, 0.18]. However, neither the fixed effect of gender of AI voice, $B = -0.08$, $SE = 0.08$, $t(6122) = -1.04$, $p = .337$, 95% CI [-0.24, 0.08], nor the interaction were significant, $B = -0.01$, $SE = 0.04$, $t(1514) = -0.18$, $p = .858$, 95% CI [-0.08, 0.07], indicating that the effect of credibility did not differ across gender of AI voice.

Participant gender as a moderator

I next conducted an exploratory analysis examining whether participant gender impacted the strength of the credibility effect. For the purpose of this analysis, I excluded participants who identified their gender as non-binary (“Other” gender; $N = 8$) or did not report their gender ($N = 4$). The fixed effects were stereotypicality (effect coded as counterstereotypical = -1, stereotypical = 1), participant gender (effect coded as woman = -1, man = 1), and their interaction. The fixed effect of stereotypicality was significant, $B = 0.13$, $SE = 0.24$, $t(2352.41) = 5.21$, $p < .001$, 95% CI [0.08, 0.17]. However, neither the fixed effect of participant gender, $B = -0.02$, $SE = 0.07$, $t(465.03) = -0.36$, $p = .719$, 95% CI [-0.16, 0.11], nor the interaction were significant, $B = -0.02$, $SE = 0.02$, $t(2353.53) = -0.66$, $p = .510$, 95% CI [-0.06, 0.03], indicating that the effect of credibility did not differ across participant gender.

Frequency of AI usage as a moderator

I next conducted an exploratory analysis examining whether frequency of prior AU usage impacted the strength of the credibility effect. The fixed effects were stereotypicality (effect coded as counterstereotypical = -1, stereotypical = 1), AI usage (grand mean centered), and their interaction. The fixed effect of stereotypicality was significant, $B = 0.12$, $SE = 0.03$, $t(1959.18) = 4.46$, $p < .001$, 95% CI [0.07, 0.17]. The fixed effect of frequency of AI usage was

also significant, $B = 0.13$, $SE = 0.05$, $t(386.95) = 2.84$, $p = .005$, 95% CI [0.04, 0.22], indicating that people with more frequent AI usage perceived the AI as more credible. However, the interaction was not significant, $B = -0.01$, $SE = 0.02$, $t(1954.45) = -0.99$, $p = .324$, 95% CI [-0.05, 0.02], indicating that the effect of credibility did not differ across participants' frequency of AI usage.

Hypothesis guessing as a moderator

Lastly, to rule out the possibility that demand effects were responsible for producing the observed effects, I conducted exploratory analyses examining whether guessing the hypothesis impacted results for the credibility effect. Two independent raters coded all open-ended responses to the question "What was this study about?". Open-ended responses were categorized into three different categories: 1 = completely irrelevant (e.g., "I don't know"; $n = 126$), 2 = mentioned something about study (e.g., "How people think of AI voices"; $n = 272$), 3 = guessed hypothesis (e.g., "How different gender and pitches in voicing affects how intelligent we think the AI is."; $n = 40$). Interrater reliability was good, $\kappa = .75$.

To examine whether guessing the hypothesis modulates results, I conducted a model in which I specified the fixed effects as stereotypicality (effect coded as counterstereotypical = -1, stereotypical = 1), hypothesis guessing contrast 1 (dummy coded as completely irrelevant = 1, mentioned something about study = 0, and guessed hypothesis = 0), hypothesis guessing contrast 2 (dummy coded as completely irrelevant = 0, mentioned something about study = 1, and guessed hypothesis = 0), and the interactions between stereotypicality and each dummy variable.

The interaction between hypothesis guessing contrast 1 and stereotypicality was significant, $B = -0.30$, $SE = 0.10$, $t(2208.91) = -3.12$, $p = .002$, 95% CI [-0.49, -0.11], indicating that the strength of the credibility effect differed between people mentioning something

irrelevant and people guessing the hypothesis. The interaction between hypothesis guessing contrast 2 and stereotypicality was also significant, $B = -0.29$, $SE = 0.09$, $t(2208.91) = -3.21$, $p = .001$, 95% CI $[-0.46, -0.11]$, indicating that the strength of the credibility effect differed between people mentioning something about the study and people guessing the hypothesis.

I also examined whether the effect of credibility differed between people who mentioned something irrelevant and people who mentioned something about the study. I conducted a model in which I specified the fixed effects as stereotypicality (effect coded as counterstereotypical = -1, stereotypical = 1), hypothesis guessing contrast 1 (dummy coded as completely irrelevant = 1, mentioned something about study = 0, and guessed hypothesis = 0), hypothesis guessing contrast 2 (dummy coded as completely irrelevant = 0, mentioned something about study = 0, and guessed hypothesis = 1), and the interactions between stereotypicality and each dummy variable. The strength of the credibility effect did not differ between people mentioning something irrelevant and people mentioning something about the study, $B = -0.01$, $SE = 0.06$, $t(2205.76) = -0.22$, $p = .827$, 95% CI $[-0.12, 0.10]$.

The effect of credibility was significant and in the predicted direction for people who guessed the hypothesis, $B = 0.39$, $SE = 0.08$, $t(2209.81) = 4.71$, $p < .001$, 95% CI $[0.23, 0.56]$, for people who mentioned something about the study, $B = 0.11$, $SE = 0.03$, $t(2202.37) = 3.34$, $p < .001$, 95% CI $[0.04, 0.17]$, and for people who mentioned something irrelevant, $B = 0.09$, $SE = 0.05$, $t(2208) = 2.00$, $p = .046$, 95% CI $[0.002, 0.19]$. These findings indicate that the effect of credibility varied depending on whether people guessed the hypothesis or not. Importantly, however, the effect was significant across *all levels* of “hypothesis guessing”, indicating that demand effects are unlikely to explain the emergence of the credibility effect.

Study 3 Summary

The results of Study 3 indicate that participants perceived an AI as more likely to provide the correct answer to a gender domain question that stereotypically matched the gender of the AI voice. This effect did not vary across gender of AI voice, participant gender, or frequency of AI usage. The effect of credibility did vary in strength depending on people's intuition of the hypothesis, but was significant among all levels of guessing the hypothesis.

Study 4

In Study 4, I sought to directly examine whether perceptions of credibility contribute to explaining the relation between gender of question domain and AI preference. Study 3 established that people perceive stereotypical (versus counterstereotypical) voice-based AI as more credible. However, in Study 3, I did not directly test whether credibility functions as a mechanism linking gender of question domain and AI voice preference. In Study 4, I predicted that (1) people would prefer gender stereotypicality in AI (replicating Studies 1-2), (2) people would perceive AI voice that stereotypically matches the gender of question domain as more credible (replicating Study 3), and (3) perceptions of credibility would in part explain the relation between gender of question domain and AI preference. The pre-registration plan for the hypotheses, methods, and analyses can be found here:

https://osf.io/m5tah/?view_only=dd32c3c6449b4c44bc508ae48288b0d4

Participants

851 participants (505 women, 338 men, 7 “Other” gender, 1 “Prefer not to answer”; $M_{age} = 41.50$; range = 18-84) were recruited from Mturk using the CloudResearch Approved Participant List. 40 additional participants completed the study, but were excluded from analyses for failing an attention check ($N = 35$), failing a soundcheck ($N = 3$), or not agreeing to have

their data included ($N = 6$).⁵ I conducted a Monte Carlo Simulation in Mplus to determine adequate sample size for this study (Muthén & Muthén, 1998-2017). I specified the relation between gender of question domain and preferred AI as $B = 0.57$ (based on Study 1), between gender of question domain and selection of AI as credible as $B = 0.13$ (based on Study 3), and between selection of AI as credible and preferred AI as $B = 0.65$. I also estimated the power for the product of .13 and .65 (the indirect effect), and then selected the sample size that was needed to detect the smallest effect (i.e., the indirect effect). A sample size of 550 participants is needed to possess at least 80% power to detect the smallest effect in the simulation (i.e., the indirect effect). I expected approximately 10% of participants to fail an attention check. To amplify power and account for attention checks, I planned to collect more participants than deemed needed by the power analysis. Thus, I aimed to collect 900 participants.

Procedure

After providing consent, participants completed the same sound check as in previous studies. Participants then performed the preference and credibility task described below. After completing all four trials of the task, participants were asked to complete questions about frequency of AI use, a demographic questionnaire, an attention check, questions about the purpose of the study, and a debrief. Because the study involved deception, participants were asked to re consent at the end of the study to have their data included in analyses.

Preference and credibility task. Participants were randomly assigned to complete 4 trials. In each trial, participants were randomly presented with 1 of 4 feminine-voiced and 1 of 4 masculine-voiced AI saying 1 of 3 gender-neutral sentences selected at random. Each AI was

⁵ The number of excluded participants does not add to 100%, because some participants failed multiple exclusion criteria.

first presented on a separate Qualtrics page (see Figure 7). Participants were shown a button (e.g., “AI 1”), and a vocal recording automatically played.

Next, both AI were displayed simultaneously on a page and participants were able to replay the vocal recordings. On this page, participants were also randomly presented with a domain specific question. In each trial, participants were randomly assigned to see either 1 of 3 possible masculine questions or 1 of 3 possible feminine questions. The order in which these questions were presented was randomized. Participants were then asked “Which AI do you think is more likely to correctly answer this question?” Participants chose between the two AI presented. On the next and final page of the block, participants saw both AI and the domain specific question again. Participants were then asked “Which AI would you use to help you solve the question?” Participants chose between the two AI presented.

Additional stipulations to these trials were that (1) there was no repetition in the voices presented to participants, and (2) there was no repetition in the question presented to participants.

Figure 7

Example Qualtrics Block Illustrating the Credibility and Preference Task

Page 1 Please listen to a voice-based AI say an example sentence. You can listen to the AI again by clicking on the button below.

AI 1

→

Page 2 Please listen to a voice-based AI say an example sentence. You can listen to the AI again by clicking on the button below.

AI 2

→

Page 3 Please click on each button below to hear two different voice-based AI.

AI 1 AI 2

Please read the following question:

Which baseball team won the world series in 1992?

Which AI do you think is more likely to correctly answer this question?

AI 1 AI 2

→

Page 4 Please click on each button below to hear two different voice-based AI.

AI 1 AI 2

Please read the following question:

Which baseball team won the world series in 1992?

Which AI would you use to help you solve the question?

AI 1 AI 2

→

Results

Analytic strategy

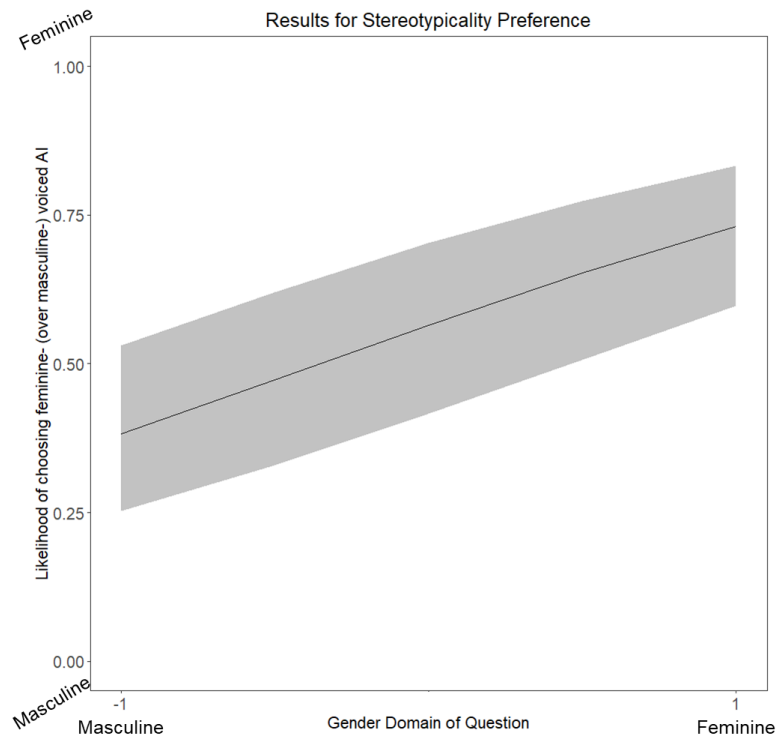
In all analyses, alpha was set to .05. For all models, I used the glmer function from the lme4 package (Bates et al., 2015) in R to account for nonindependence in judgments (Fitzmaurice et al., 2012). In all models, I specified random intercepts of participant, feminine AI voice, masculine AI voice, question, gender-neutral sentence said by a feminine voice, and gender-neutral sentence said by a masculine voice to account for between-participant and between-stimuli variance (Judd et al., 2012).

Do participants prefer gender stereotypicality (versus counterstereotypicality) in AI?

I first conducted a confirmatory analysis examining whether the stereotypicality effect emerges. I specified the fixed effect of gender domain of question (effect coded as masculine = -1, feminine = 1) as the predictor variable, and preferred AI as the dependent variable (coded as masculine = 0, feminine = 1). The main effect of gender domain was significant, $B = 0.74$, $SE = 0.04$, Wald $\chi^2 = 375.86$, $p < .001$, 95% CI [0.66, 0.81], odds ratio = 2.09, indicating that participants were more likely to choose a feminine- (versus masculine-) voiced AI when asked to answer a question in a feminine (versus masculine) domain (see Figure 8). This finding replicates Studies 1-2, and supports that people prefer gender stereotypicality (versus counterstereotypicality) in voiced-based AI.

Figure 8

Likelihood of Choosing Feminine- (Over Masculine-) Voiced AI as a Function of the Gender Domain of the Question (Study 4).



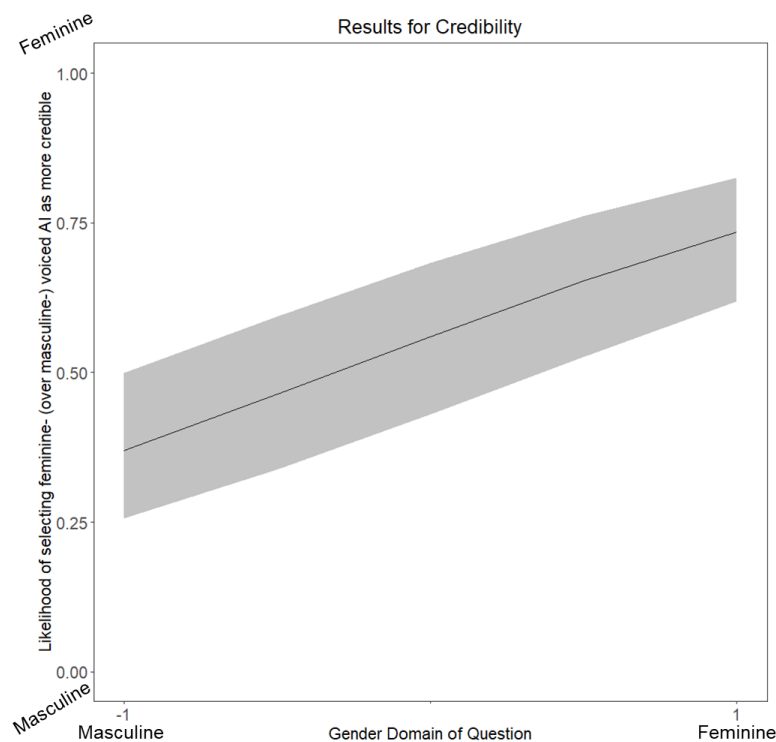
Do participants perceive stereotypical (versus counterstereotypical) AI as more credible?

I next conducted a confirmatory analysis examining whether participants show an increased tendency to select feminine-voiced (over masculine-voiced) AI as more likely to correctly answer a question in a feminine (versus masculine) domain. I specified the fixed effect of gender domain of question (effect coded as masculine = -1, feminine = 1) as the predictor variable, and selection of AI as more credible as the dependent variable (coded as masculine = 0, feminine = 1). The main effect of gender domain was significant, $B = 0.78$, $SE = 0.05$, Wald $\chi^2 = 293.51$, $p < .001$, 95% CI [0.69, 0.87], odds ratio = 2.17, indicating that participants selected a

feminine-voiced (over masculine-voiced) AI as more likely to correctly answer a question in a feminine (versus masculine) domain (see Figure 9). This finding supports that people are more likely to believe an AI voice that stereotypically matches the gender of question domain to be more credible.

Figure 9

Likelihood of Selecting Feminine- (Over Masculine-) Voiced AI as More Credible as a Function of the Gender Domain of the Question (Study 4).



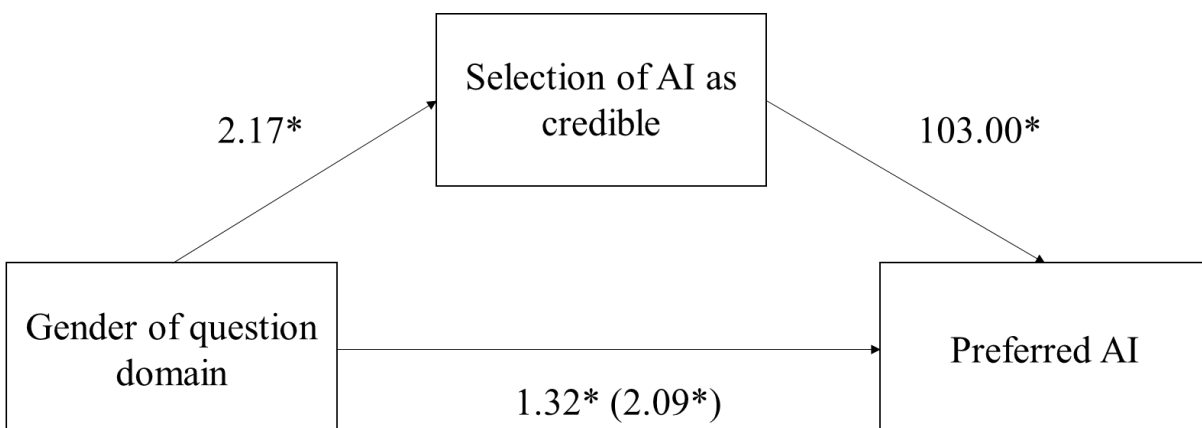
Is the relation between gender of question domain and AI preference in part explained by credibility of the AI?

I conducted a 1-1-1 mediation model with fixed slopes using multilevel structural equation modeling (MSEM) in Mplus. I specified gender of question domain (effect coded as

masculine = -1, feminine = 1) as the exogenous variable, selection of AI as more credible (coded as masculine = 0, feminine = 1) as the mediator variable, and preferred AI (coded as masculine = 0, feminine = 1) as the dependent variable. I specified participant as the cluster variable. The indirect was significant, 95% CI [0.27, 3.05] (see Figure 10), indicating that greater likelihood of selecting a feminine-voiced (versus masculine-voiced) AI as more credible in a feminine (versus masculine) domain in part explained why people were more likely to choose a feminine-voiced (over masculine-voiced) AI in a feminine question domain. In other words, the relation between gender of question domain and gender of AI voice preference was in part explained by selection of AI as more credible.

Figure 10

Model in which Gender of Question Domain is Associated with Gender of Preferred AI Through Gender of AI Selected as More Credible (Study 4).



Note. Values are odds ratios. Value in bracket represents unadjusted effect. Value outside of brackets represents effects including all variables in the model. $*p < .001$

Exploratory Analyses

In the next section, I report exploratory analyses examining whether the stereotypicality and credibility effects vary as a function of the gender domain of question, participant gender, and frequency of prior AI usage. To address potential demand characteristics of the study's design, I also assessed whether correctly guessing the hypothesis would modulate results.

Variation across gender of question domain (preferred AI as dependent variable)

I conducted exploratory analyses examining whether the effect of preferred AI emerges in each gender domain of the question. To examine preferences within the masculine and feminine question conditions separately, I conducted generalized linear mixed models with preferred AI (coded as masculine = 0, feminine = 1) as the dependent variable.

Feminine domain. To test the effect of preference within the feminine domain, the fixed effect in the model was gender domain of the question (dummy coded as feminine = 0, masculine = 1). The intercept was significant and positive, $B = 1.00$, $SE = 0.31$, Wald $\chi^2 = 10.50$, $p = .001$, 95% CI [0.39, 1.60], odds ratio = 2.72, indicating that the likelihood of choosing a feminine- (versus masculine-) voiced AI was higher when presented with feminine domain questions.

Masculine domain. To test the effect of preference within the masculine domain, the fixed effect in the model was gender domain of the question (dummy coded as masculine = 0, feminine = 1). The intercept was not significant, $B = -0.48$, $SE = 0.31$, Wald $\chi^2 = 2.42$, $p = .120$, 95% CI [-1.08, 0.12], odds ratio = 0.62, indicating that the likelihood of choosing a feminine- (versus masculine-) voiced AI did not significantly differ for masculine domain questions.

Variation across gender of question domain (credibility as dependent variable)

I also conducted exploratory analyses examining whether the effect of AI credibility emerges in each gender domain of the question. To examine selection of AI as more credible within the masculine and feminine question conditions separately, I conducted generalized linear mixed models with selection of AI as more credible (coded as masculine = 0, feminine = 1) as the dependent variable.

Feminine domain. To test the effect of credibility within the feminine domain, the fixed effect in the model was gender domain of the question (dummy coded as feminine = 0, masculine = 1). The intercept was significant and positive, $B = 1.02$, $SE = 0.27$, Wald $\chi^2 = 14.00$, $p < .001$, 95% CI [0.49, 1.55], odds ratio = 2.77, indicating that the likelihood is higher of selecting a feminine- (versus masculine-) voiced AI as credible for feminine domain questions.

Masculine domain. To test the effect of credibility within the masculine domain, the fixed effect in the model was gender domain of the question (dummy coded as masculine = 0, feminine = 1). The intercept was significant and negative, $B = -0.54$, $SE = 0.27$, Wald $\chi^2 = 3.89$, $p = .049$, 95% CI [-1.07, -0.003], odds ratio = 0.59, indicating that the likelihood is higher of selecting a masculine- (versus feminine-) voiced AI as credible for masculine domain questions.

Participant gender as a moderator

I next conducted an exploratory analysis examining whether participant gender impacted the strength of the effects. For the purpose of this analysis, I excluded participants who identified their gender as non-binary (“Other” gender; $N = 7$) or did not report their gender ($N = 1$). The fixed effects were gender domain of the question (effect coded as masculine = -1, feminine = 1), participant gender (effect coded as woman = -1, man = 1), and their interaction.

Preferred AI as dependent variable. First, I conducted a model with preferred AI as the dependent variable (coded as masculine = 0, feminine = 1). The fixed effect of gender domain was significant, $B = 0.74$, $SE = 0.04$, Wald $\chi^2 = 358.61$, $p < .001$, 95% CI [0.67, 0.82], odds ratio = 2.10, indicating that the likelihood of choosing a feminine- (over masculine-) voiced AI was higher when asked to answer a question in a stereotypically feminine (versus masculine) domain. The fixed effect of participant gender was also significant, $B = 0.08$, $SE = 0.04$, Wald $\chi^2 = 3.96$, $p = .047$, 95% CI [0.001, 0.15], odds ratio = 1.08, indicating that the likelihood of choosing a feminine- (over masculine-) voiced AI was higher for men than for women. The interaction was not significant, $B = 0.02$, $SE = 0.04$, Wald $\chi^2 = 0.22$, $p = .635$, 95% CI [-0.06, 0.09], odds ratio = 1.02, indicating that the likelihood of choosing a feminine- (over masculine-) voiced AI based on gender domain did not differ based on participant gender.

Credibility as dependent variable. I also conducted an exploratory analysis examining whether participant gender impacted the strength of the credibility effect. The model was specified as noted above, and the dependent variable was the selection of AI as more credible (coded as masculine = 0, feminine = 1). The fixed effect of gender domain was significant, $B = 0.77$, $SE = 0.05$, Wald $\chi^2 = 276.82$, $p < .001$, 95% CI [0.68, 0.86], odds ratio = 2.15, indicating that the likelihood of selecting a feminine- (over masculine-) voiced AI as credible was higher in a stereotypically feminine (versus masculine) domain. Neither the fixed effect of participant gender, $B = 0.03$, $SE = 0.04$, Wald $\chi^2 = 0.43$, $p = .514$, 95% CI [-0.05, 0.10], odds ratio = 1.03, nor the interaction were significant, $B = -0.06$, $SE = 0.04$, Wald $\chi^2 = 2.19$, $p = .139$, 95% CI [-0.13, 0.02], odds ratio = 0.94. These findings indicate that the likelihood of selecting a feminine- (over masculine-) voiced AI as credible based on gender of question domain did not differ based on participant gender.

Frequency of AI usage as a moderator

I next conducted an exploratory analysis examining whether frequency of prior AI usage impacted the strength of the effects. The fixed effects were gender domain of question (effect coded as masculine = -1, feminine = 1), AI usage (grand mean centered), and their interaction.

AI preference as dependent variable. First, I conducted a model with preferred AI as the dependent variable (coded as masculine = 0, feminine = 1). The fixed effect of gender domain was significant, $B = 0.72$, $SE = 0.05$, Wald $\chi^2 = 229.64$, $p < .001$, 95% CI [0.63, 0.82], odds ratio = 2.06, indicating that the likelihood of choosing a feminine- (over masculine-) voiced AI was higher when asked to answer a question in a stereotypically feminine (versus masculine) domain. Neither the main effect of AI usage, $B = -0.01$, $SE = 0.03$, Wald $\chi^2 = 0.31$, $p = .579$, 95% CI [-0.07, 0.04], odds ratio = 0.99, nor the interaction between gender domain and AI usage were significant, $B = -0.02$, $SE = 0.03$, Wald $\chi^2 = 0.63$, $p = .427$, 95% CI [-0.07, 0.03], odds ratio = 0.98. These findings indicate that the preference for gender stereotypicality in AI did not vary based on frequency of prior AI usage.

Credibility as dependent variable. I then conducted the same exploratory analysis as noted above to examine whether frequency of prior AI usage impacted the strength of the credibility effect. The dependent variable was the selection of AI as more credible (coded as masculine = 0, feminine = 1). The fixed effect of gender domain was significant, $B = 0.77$, $SE = 0.05$, Wald $\chi^2 = 196.73$, $p < .001$, 95% CI [0.66, 0.88], odds ratio = 2.16, indicating that the likelihood of selecting a feminine- (over masculine-) voiced AI as more credible was higher in a stereotypically feminine (versus masculine) domain. Neither the main effect of AI usage, $B = -0.02$, $SE = 0.03$, Wald $\chi^2 = 0.45$, $p = .502$, 95% CI [-0.07, 0.03], odds ratio = 0.98, nor the interaction between gender domain and AI usage were significant, $B = -0.04$, $SE = 0.03$, Wald χ^2

= 2.75, $p = .097$, 95% CI [-0.10, 0.01], odds ratio = 0.96. These findings indicate that the selection of an AI voice that stereotypically matches the gender of question domain to be more credible did not vary based on frequency of prior AI usage.

Hypothesis guessing as a moderator

Lastly, to rule out the possibility that demand effects were responsible for producing the observed effects, I conducted exploratory analyses examining whether guessing the hypothesis impacted results for the gender stereotypicality and credibility effects. Two independent raters coded all open-ended responses to the question “What was this study about?”. Open-ended responses were categorized into three different categories for each of the hypotheses.⁶

Stereotypicality hypothesis. For the stereotypicality hypothesis, responses were categorized into the following three categories: 1 = completely irrelevant (e.g., “I’m not sure”; $n = 289$), 2 = mentioned something about study (e.g., “Judging AI voices”; $n = 533$), 3 = guessed hypothesis (e.g., “Whether feminine questions make you want a feminine AI voice [...]”; $n = 29$). Interrater reliability was good, $\kappa = .79$.

To examine whether guessing the hypothesis modulates results for the stereotypicality effect, I conducted a model in which I specified the fixed effects as gender of question domain (effect coded as masculine = -1, feminine = 1), hypothesis guessing contrast 1 (dummy coded as completely irrelevant = 1, mentioned something about study = 0, and guessed hypothesis = 0), hypothesis guessing contrast 2 (dummy coded as completely irrelevant = 0, mentioned something about study = 1, and guessed hypothesis = 0), and the interactions between gender

⁶ Raters also coded responses for the mediation hypothesis ($\kappa = .88$). However, no participant guessed this hypothesis, and so it is not included in analyses of demand effects.

domain of question and each dummy variable. The dependent variable was preferred AI (coded as masculine = 0, feminine = 1).

The interaction between hypothesis guessing contrast 1 and gender of question domain was significant, $B = -0.85$, $SE = 0.25$, Wald $\chi^2 = 11.80$, $p < .001$, 95% CI $[-1.33, -0.36]$, odds ratio = 0.43, indicating that the strength of the stereotypicality effect differed between people mentioning something irrelevant and people guessing the hypothesis. The interaction between hypothesis guessing contrast 2 and gender of question domain was also significant, $B = -0.69$, $SE = 0.24$, Wald $\chi^2 = 8.00$, $p = .005$, 95% CI $[-1.16, -0.21]$ odds ratio = 0.50, indicating that the strength of the stereotypicality effect differed between people mentioning something about the study and people guessing the hypothesis.

I also examined whether the effect of stereotypicality differed between people who mentioned something irrelevant and people who mentioned something about the study. I conducted a model in which I specified the fixed effects as gender of question domain (effect coded as masculine = -1, feminine = 1), hypothesis guessing contrast 1 (dummy coded as completely irrelevant = 1, mentioned something about study = 0, and guessed hypothesis = 0), hypothesis guessing contrast 2 (dummy coded as completely irrelevant = 0, mentioned something about study = 0, and guessed hypothesis = 1), and the interactions between gender domain of question and each dummy variable. The strength of the stereotypicality effect differed between people mentioning something irrelevant and people mentioning something about the study, $B = -0.16$, $SE = 0.08$, Wald $\chi^2 = 4.04$, $p = .045$, 95% CI $[-0.32, -0.004]$, odds ratio = 0.85.

The effect of stereotypicality was significant and in the predicted direction for people who guessed the hypothesis, $B = 1.46$, $SE = 0.24$, Wald $\chi^2 = 37.60$, $p < .001$, 95% CI $[0.99, 1.93]$, odds ratio = 4.31, for people who mentioned something about the study, $B = 0.78$, $SE =$

0.05, Wald $\chi^2 = 256.38$, $p < .001$, 95% CI [0.68, 0.87], odds ratio = 2.17, and for people who mentioned something irrelevant, $B = 0.62$, $SE = 0.06$, Wald $\chi^2 = 94.01$, $p < .001$, 95% CI [0.49, 0.74], odds ratio = 1.17. These findings indicate that although the effect of stereotypicality varied across levels of “hypothesis guessing”, it still emerged regardless of whether people guessed the hypothesis or not.

Credibility hypothesis. For the credibility hypothesis, responses were categorized into the following three categories: 1 = completely irrelevant (e.g., “I’m not sure”; $n = 289$), 2 = mentioned something about study (e.g., “Judging AI voices”; $n = 537$), 3 = guessed hypothesis (e.g., “Whether gender bias plays a role in determining who is more knowledgeable.”; $n = 25$). Interrater reliability was good, $\kappa = .72$.

To examine whether guessing the hypothesis modulates results for the credibility effect, I conducted a model in which I specified the fixed effects as gender of question domain (effect coded as masculine = -1, feminine = 1), hypothesis guessing contrast 1 (dummy coded as completely irrelevant = 1, mentioned something about study = 0, and guessed hypothesis = 0), hypothesis guessing contrast 2 (dummy coded as completely irrelevant = 0, mentioned something about study = 1, and guessed hypothesis = 0), and the interactions between gender domain of question and each dummy variable. The dependent variable was the selection of AI as more credible (coded as masculine = 0, feminine = 1).

The interaction between hypothesis guessing contrast 1 and stereotypicality was significant, $B = -0.65$, $SE = 0.27$, Wald $\chi^2 = 5.59$, $p = .018$, 95% CI [-1.18, -0.11], odds ratio = 0.52, indicating that the strength of the credibility effect differed between people mentioning something irrelevant and people guessing the hypothesis. The interaction between hypothesis guessing contrast 2 and stereotypicality was not significant, $B = -0.36$, $SE = 0.27$, Wald $\chi^2 =$

1.73, $p = .188$, 95% CI [-0.89, 0.17], odds ratio = 0.70, indicating that the strength of the credibility effect did not differ between people mentioning something about the study and people guessing the hypothesis.

I also examined whether the effect of credibility differed between people who mentioned something irrelevant and people who mentioned something about the study. I conducted a model in which I specified the fixed effects as gender of question domain (effect coded as masculine = -1, feminine = 1), hypothesis guessing contrast 1 (dummy coded as completely irrelevant = 1, mentioned something about study = 0, and guessed hypothesis = 0), hypothesis guessing contrast 2 (dummy coded as completely irrelevant = 0, mentioned something about study = 0, and guessed hypothesis = 1), and the interactions between gender domain of question and each dummy variable. The strength of the credibility effect differed between people mentioning something irrelevant and people mentioning something about the study, $B = -0.29$, $SE = 0.08$, Wald $\chi^2 = 13.45$, $p < .001$, 95% CI [-0.45, -0.14], odds ratio = 0.75.

The effect of credibility was significant and in the predicted direction for people who guessed the hypothesis, $B = 1.23$, $SE = 0.27$, Wald $\chi^2 = 21.10$, $p < .001$, 95% CI [0.70, 1.75], odds ratio = 3.41, for people who mentioned something about the study, $B = 0.87$, $SE = 0.06$, Wald $\chi^2 = 247.65$, $p < .001$, 95% CI [0.76, 0.98], odds ratio = 2.39, and for people who mentioned something irrelevant, $B = 0.58$, $SE = 0.07$, Wald $\chi^2 = 72.37$, $p < .001$, 95% CI [0.45, 0.72], odds ratio = 1.79.

Overall, these findings indicate that the effects of stereotypicality and credibility varied across levels of “hypothesis guessing”. Importantly, however, the effects were significant across *all levels* of “hypothesis guessing”, indicating that demand effects are unlikely to explain the emergence of the stereotypicality or credibility effects.

Study 4 Summary

The results of Study 4 indicate that people prefer stereotypicality (over counterstereotypicality) in voice-based AI and were more likely to select an AI voice that stereotypically matches the gender of question domain to be more credible. Importantly, the findings provide support for the explanatory role of perceived credibility. Specifically, perceptions of credibility in part explained the preferred stereotypical relation between gender of question domain and AI preference. The effect of preference varied in strength across feminine and masculine domains, such that people were more likely to prefer a feminine- (over masculine-) voiced AI in a feminine domain, but were not more likely to prefer a masculine- (over feminine-) voiced AI in a masculine domain. Importantly, the effect of preference did not vary as a function of participant gender or frequency of AI use. The effect of credibility did not vary as a function of gender of question domain, participant gender, or frequency of AI use. The stereotypicality and credibility effects varied in strength depending on people's intuition of the hypotheses, but were significant among all levels of guessing the hypothesis.

CHAPTER 5: GENERAL DISCUSSION

In the present research, I addressed the question of whether people prefer for voice-based AI to align with gender stereotypes, and examined perceived credibility as a potential mechanism for this preference. Across Studies 1 and 2, I found that people preferred stereotypical (over counterstereotypical and androgynous) voice-based AI. In Studies 3 and 4, I found that people perceived an AI as more credible when the gender of the AI voice was a stereotypical (versus counterstereotypical) match to the gender of the question. I also found that the selection of a stereotypical AI as more credible in part explained people's preference for gender stereotypicality. The present research is uniquely suited to examine people's reliance on gender as an organizing principle when interacting with AI, as AI is a non-human entity without biological sex. These findings hold implications for understanding people's preferences for stereotypicality in non-human entities, the role of androgyny in a context removed from societal gender norms, variation across gender of question domain for the credibility effect, and limitations and implications of this work.

Preferences for Gender Stereotypicality in AI

Prior research highlights that people categorize and stereotype to make sense of non-human entities (Auster & Mansbach, 2012; Jung et al., 2014; Weisgram et al., 2014; Wilkie & Bodenhausen, 2012, 2015). However, whether these social-cognitive processes extend into a preference for gender stereotypicality in non-human entities was unknown. The present research addressed this question and provides support for a systematic preference for stereotypicality in AI. Despite AI representing a blank slate without biological sex and a context not governed by the same norms of social conduct as in human interactions, people preferred stereotypicality over counterstereotypicality and androgyny in voice-based AI. Thus, people not only categorize non-

human entities according to their perceived group membership but also prefer stereotypical congruence between a gendered domain and AI voice. Preference for gender stereotypicality in AI exists primarily due to the application of cultural stereotypes to non-human entities. People seemingly do not inhibit or censor their preferences in AI to appear egalitarian within this context. As such, AI might be used to more freely examine cognitive processes that social desirability typically affects. Overall, these findings contribute to our understanding of how people use gender stereotypes outside prevailing human gender relations, which holds implications for how AI might be used to make sense of the social world, as well as create and reinforce social reality.

The present research utilized AI to examine preferences for gender stereotypicality as it provides a unique context removed from social and non-social factors that usually govern the expression of bias. An important question that arises is whether the results simply replicate findings from human targets or whether there is something qualitatively different about examining social cognitive processes in AI. I argue that while results from the present studies largely replicate findings among human targets, there are important differences suggesting that AI may represent a unique social context. For instance, the effect of gender stereotypicality is stronger in the feminine compared to the masculine domain in AI. Specifically, the likelihood of choosing a feminine or masculine AI did not differ in the masculine domain. This result is somewhat counterintuitive and surprising. Past research examining preferences for stereotypicality in human targets suggests that fulfillment of stereotypical social roles tends to be positively evaluated. However, failure to conform to prescriptive gender norms often results in negative reactions and evaluations from others, as well as social and economic penalties (i.e., Backlash effect; Rudman, 1998; Mahalik, 2000; Rudman & Fairchild, 2004; Rudman & Glick,

2001). Thus, incongruence between characteristics of a human group and requirements of a social role is evaluated negatively (e.g., women in leadership roles or men in caregiver roles). For instance, displays of counterstereotypical traits and behavior in men are associated with decreased likeability (Kite & Deaux, 1987; Rudman, 1998; Rudman & Glick, 2001), and decreased perceptions of competency (Rudman, 1998; Rudman & Glick, 1999, 2001). In addition, manhood or masculinity is seen as hard to attain but easy to lose, thus making it especially precarious (Vandello et al., 2008; Vandello & Bosson, 2013).

Taken together, predictions derived from literature on human targets would suggest a stronger preference for stereotypicality in a masculine domain. However, I observed the exact opposite. Thus, previous research examining evaluations based on gender stereotypicality in humans contrasts with the present findings. A potential explanation for the diverging findings could be that the present studies' task of asking an AI to provide an answer to a question provokes stereotypical beliefs in people's minds (e.g., viewing question answering as a feminine task). Stereotypical beliefs associated with the task itself may override preferences in a masculine domain. Specifically, stereotypes about women as helpful, friendly, and supportive (e.g., Bem, 1974; Eagly & Mladinic, 1989) make them a potentially valid option in a question-answering task regardless of the gendered nature of the domain. In the present research, we choose a question-answering task because it is a common way to utilize AI in everyday life. Future research could examine whether changing the task to be more or less feminine/masculine affects people's preferences for stereotypicality in AI.

Another way to look at the present results is through examining the degree to which people who chose a feminine- (or masculine-) voiced AI did so within a stereotypical or counterstereotypical domain. Exploratory analyses suggest that the likelihood of making a

stereotypical (versus counterstereotypical) domain choice is lower for feminine- (versus masculine-) voiced AI (Study 1: $B = -0.18$, $SE = 0.06$, Wald $\chi^2 = 9.17$, $p = .003$; Study 2: $B = -0.46$, $SE = 0.11$, Wald $\chi^2 = 16.20$, $p < .001$; Study 4 stereotypicality effect: $B = -0.15$, $SE = 0.04$, Wald $\chi^2 = 13.99$, $p < .001$; Study 4 credibility effect: $B = -0.15$, $SE = 0.04$, Wald $\chi^2 = 13.04$, $p < .001$). This pattern of result suggests that people are more likely to select a feminine-voiced AI to answer any question, whereas they perceive the masculine voice to have more domain specific applicability. In addition, as discussed above, expectations about women's cooperation and willingness to help may make people more inclined to choose a feminine-voiced AI for either a masculine or feminine question. However, it is also important to note that people were not more likely to choose a feminine- (over masculine-) voiced AI overall.

Interestingly, findings in the present studies largely were not moderated by participant sex/gender or frequency of AI use. The effect of gender stereotypicality only varied as a function of participant sex in Study 1. The preference for gender stereotypicality emerged among both male and female participants, providing support for the main hypothesis, but female participants showed a stronger preference than male participants. These findings are largely in line with past research, which finds that both men and women prefer gender stereotypicality and evaluate deviations from it negatively (Rudman, 1998; Mahalik, 2000; Rudman & Fairchild, 2004; Rudman & Glick, 2001). I also examined whether prior familiarity with AI moderated the effects. Across studies, effects were not influenced by people's frequency of AI use. These findings suggest that past experiences with AI did not affect people's preferences.

Devaluation of Androgyny in AI

A wealth of research has examined preferences for stereotypicality over counterstereotypicality in humans (for a review, see Morgenroth & Ryan, 2018), and some

research from the 1970s and early 1980s primarily conducted by Sandra Bem and Janet Spence directly examined issues surrounding the perception and expression of androgyny. However, only recently has research begun to re-examine people's perceptions toward androgyny (Atwood & Axt, 2021; Morgenroth & Ryan, 2018). In the present research, I examined whether providing the option of a non-stereotypically voiced (i.e., androgynous) AI would affect people's preferences. Some past findings point toward the possibility that androgynous individuals are evaluated more positively than gender stereotypical people (Basow & Howe, 1987; Hester et al., 2021). In addition, a context removed from biological sex, such as AI, may allow for an androgyny preference to emerge. However, findings from Study 2 did not support this idea. People strongly devalued androgyny in AI, instead showing a stronger preference for gender stereotypicality and counterstereotypicality. These findings are in line with literature on perceptual fluency, which posits that the harder it is to process stimuli, the less favorable they are evaluated (Winkielman et al., 2006).

In the present research, androgyny was operationalized as voices that are not perceived as prototypical or stereotypical of men or women. Thus, androgynous voices are ambiguous and not easily categorized, likely resulting in processing difficulties. These findings are in line with extant findings that people dislike ambiguity in social categorization. For instance, the processing of targets that disrupt gender or race binaries, such as multiracial or transgender people, is associated with difficulties in processing (Lick et al., 2015; Remedios et al., 2012). It is important to note that I only altered the androgynous voices in the present research, and not the masculine or feminine voices. Thus, an alternative explanation for people's devaluation of androgyny could be due to androgynous AI being perceived as non-human due to altering the voice and not its quality of non-stereotypicality. While the involved researchers perceived

androgynous voices as having human (versus machine-like) qualities, future research should more closely examine whether altering all voices would lead to different results.

The present findings may also hold important implications for understanding the perception of humans displaying androgynous qualities. Nonbinary people, who identify their gender outside of the man/woman binary, have become increasingly visible in society and often face severe discrimination (Wilson & Meyer, 2021). How people perceive androgyny in AI potentially provides insight into how androgyny is viewed in humans. Initial theorizing outlined an argument that removing non-human entities from societal gender relations may offer a unique opportunity to minimize the use of gender as an organizing principle and lead to less discrimination (Haraway, 1987, 1997). However, the present research suggests that even in contexts removed from societal gender relations, such as AI, people prefer stereotypicality. These findings point to the possibility that gender biases persist and are (re)constructed within non-human entities without biological sex markers.

Moreover, recent research highlights that gender is integral to being perceived as human. Martin and Mason (2022) found that the absence of gender leads to dehumanization of targets. In one of their studies, participants were asked to imagine either male, female, or genderless voice assistants. People describing a genderless (versus female or male) voice assistant were less likely to ascribe humanness. Even though it is not entirely clear what people were imagining when prompted to think of a genderless AI (and it might be different from an androgynous AI), these results still provide some important insight into why people might devalue androgyny in AI. People may dehumanize androgynous AI relative to a stereotypical or counterstereotypical AI, which would explain why people preferred androgyny least. Relatedly, prior theorizing has also described a general effect of perceiving targets closest to the bounds of humanness negatively

(i.e., uncanny valley hypothesis; Mori, 1970). The theory posits that the more humanlike an entity is perceived to be, the more difficulty people experience in discerning whether the target is a human or a humanlike object. Trouble in distinguishing between human and humanlike entities provokes feelings of eeriness and discomfort, which result in more negative evaluations of the humanlike object. These negative evaluations of near-humanlike entities stand in contrast to an otherwise positive association between human likeness and affect. Thus, androgynous targets are potentially devalued in part because they are perceived as less human (or not human enough). More closely examining mechanisms associated with people's devaluation of androgyny would be a fruitful avenue for future research.

In addition, this line of reasoning also poses questions about the difference between *androgyny* and *genderlessness*. In the present research, I operationalized androgyny through voices that were neither stereotypical nor prototypical of women or men. Past research has described androgyny as scoring high on both femininity and masculinity, whereas “undifferentiated” individuals are defined through low scores on both femininity and masculinity (Bem, 1974). Scores on masculinity and femininity were also related to mental health, such that undifferentiated individuals scored lowest on mental health (Spence, 1993). It is unclear whether androgyny, undifferentiated, nonbinary, and genderless present four theoretically distinct constructs. For instance, it remains largely unexamined whether genderless and undifferentiated are the same or different. Future research could more closely attempt to tease apart these constructs and then examine whether members of these groups differ in terms of other constructs (e.g., mental health outcomes).

A related and yet unanswered question concerns individual differences in preferences for gender stereotypicality over counterstereotypicality and androgyny. Interestingly, even though

overall people devalued androgyny and showed stronger preferences for stereotypicality and counterstereotypicality, in about 8% of all trials in Study 2 participants chose an androgynous- (over feminine- or masculine-) voiced AI. Given that processing of ambiguous stimuli is more cognitively taxing, these choices are surprising. Underlying individual motivations and differences guiding people's preferences for stereotypicality in AI remain unclear. People may override their initial preference to align their preference with personal egalitarian goals (e.g., Plant & Devine, 1998). In contrast, people preferring androgyny may hold more positive attitudes overall towards people disrupting the gender binary (Stern, 2022). Future research could more closely examine whether individual differences, such as need for structure, egalitarian values, or prejudicial attitudes, are associated with preferring gender stereotypicality, counterstereotypicality, or androgyny in AI.

Variation Across Gender of Question Domain (Credibility Effect)

Preferences for stereotypicality in humans are in part determined by specific behavioral expectations. In Studies 3 and 4, I examined whether people perceive a stereotypical (versus counterstereotypical) AI as more credible. I found that perceived credibility in part explained the correspondence between gender of question domain and preferred gender of AI voice. Despite AI being intelligent machines that hold the same knowledge independent of the gender of their voice, people perceived gender differences in an AI's ability to correctly answer a question in a feminine or masculine domain. These findings support the notion that people use gender as a schema when interacting with AI, which shapes their perceptions and expectations of AI in accordance with human stereotypes.

I found that the effect of credibility emerged in both feminine and masculine domains, indicating that people perceive a feminine-voiced AI as more credible in a feminine domain and

a masculine-voiced AI as more credible in a masculine domain. This pattern of findings stands in contrast to the effect of stereotypicality preference which primarily emerged in feminine but not masculine domains. Although people perceive a stereotypical match between gender of the voice and question domain in voice-based AI as more credible, this perception does not perfectly align with people's preferences in both feminine and masculine domains. Future research could investigate the underlying reasons for this discrepancy.

Limitations and Implications

The present studies are a first step to examining preferences for stereotypicality in AI. I examined this in the context of voices from White, cisgender, straight volunteers with a Midwestern accent. However, speech and use of language vary based on people's cultural backgrounds and social identities (race: Jones, 2015; Purnell et al., 1999; sexual orientation: Linville, 1998; Mack & Munson, 2012). In addition, people can discriminate speech and categorize, for instance, the speaker's race (Perrachione & Wong, 2007; Remez et al., 1997), and sexual orientation (Rieger et al., 2010; Tracy et al., 2015). Thus, differences may emerge as a function of how an AI's voice group membership is perceived. A growing body of literature has highlighted the importance of examining evaluations of people holding multiple (marginalized) identities simultaneously (e.g., Purdie-Vaughns & Eibach, 2008; Remedios & Snyder, 2015). Social categories do not shape social judgments independently and, instead, the intersection of social categories uniquely shapes judgments and evaluations (e.g., Remedios et al., 2011). For example, listeners process and are able to categorize speakers based on auditory cues from speakers with intersecting identities (Levon, 2014, 2015).

It is unclear how these evaluations based on intersecting identities would translate to preferences for stereotypicality in AI. Research highlights the primacy of gender in social

judgments (Martin & Slepian, 2021; Martin & Mason, 2022), which provides reason to believe that present findings would hold regardless of race. However, research examining perceptions at the intersection of race and gender highlights that Black women are perceived as more masculine than White women (Galinsky et al., 2013; Goff et al., 2008). Therefore, stereotypes about race and gender may result in more complex patterns of results. Future research could more closely examine whether and how results hold when examining AI voices that are perceived as being at the intersection of different social identities.

Given the increased use of AI through voice-based assistants, it remains important to consider how exposure to gendered voice-based AI may influence people's beliefs about others and society at large. Specifically, whether exposure to gender stereotypical or counterstereotypical AI disrupts people's stereotype endorsement and evaluation of people who challenge gender stereotypes stands as an open question. Initial evidence highlights that engaging in social processes with non-human entities has downstream consequences for subsequent evaluations of humans (Martin & Slepian, 2018). Thus, exposure to (counter)stereotypical AI might shape stereotype endorsement and subsequent evaluations of humans who challenge stereotypes.

Past research examining exposure to human exemplars has found that stereotypical congruence between the category of the person and the category of the occupied domain reinforces stereotypical associations (Dasgupta & Greenwald, 2001). For instance, women primed with traditional (versus non-traditional) gender roles reported greater implicit gender stereotypes (Rudman & Phelan, 2010). In contrast, exposure to counterstereotypical exemplars often reduces stereotype endorsement and the expression of (gender) bias (Dasgupta & Asgari, 2004; Lai et al., 2014). For example, individuals presented with women in leadership roles

showed decreased implicit stereotyping (Dasgupta & Asgari, 2004). Thus, exemplars might generalize to the group as a whole and affect stereotype endorsement. In line with this reasoning, exposure to counterstereotypical exemplars might decrease stereotype endorsement leading to decreased stereotype application. Thus, counterstereotypical exemplars may not only reduce stereotype endorsement but also reduce discrimination against people who challenge the stereotypes of their group. Future research could more closely examine downstream consequences of exposure to stereotypical and counterstereotypical AI.

Preference for gender stereotypicality in AI may reinforce current societal gender roles. Past research has primarily examined the application of social categories to non-human entities. In addition, present results highlight people's preferences for gender stereotypicality. However, AI are not only unique because they lack biological sex, but also are algorithms that are capable to learn from prior experiences (e.g., Collins & Moons, 2019; Gepperth & Karaoguz, 2016; Parisi et al., 2019). For instance, AI algorithms are capable of using input from the environment to improve their predictions. Given this interactive and malleable nature of AI algorithms, the application of gender to AI may hold important consequences for both humans and AI. How do human-AI interactions alter the AI itself and, subsequently, humans? The present research suggests that people specifically seek out feminine-voiced AI for feminine tasks given their assumed high credibility in feminine domains. To the extent that AI learn and become more proficient through repeated application in a domain, AI might develop or acquire something akin to human gender. This could result in a self-reinforcing cycle, which may ultimately reinforce current gender roles. Thus, through social-constructive processes AI may be altered and engrain gender stereotypes even in the absence of biological sex. Earlier theoretical conceptions highlight mechanisms through which machines may reduce essentialist understanding of

sex/gender (Haraway, 1985). In contrast, the present line of reasoning would suggest that gender essentialism may instead be maintained or potentially reinforced through gendered AI in the future.

Concluding Remarks

Gender is primary in social categorization. In the present research, I argued that people's preferences for humans similarly extend to non-human entities. I tested these ideas using AI which presents a unique space that is removed from social and non-social factors that commonly govern the expression and application of gender stereotypes to human targets. Across four studies, I found that people preferred gender stereotypicality (over counterstereotypicality and androgyny) in voice-based AI and that this preference was in part explained by people's beliefs that a stereotypical AI was more credible. These findings help to advance knowledge on people's reliance on gender as an organizing principle and contribute to our understanding of how people use gender stereotypes outside of human gender relations to create and reinforce social reality.

REFERENCES

- Abele, A. E. (2003). The dynamics of masculine-agentic and feminine-communal traits: Findings from a prospective study. *Journal of Personality and Social Psychology*, 85(4), 768–776.
<https://doi.org/10.1037/0022-3514.85.4.768>
- Allport, G. W., Clark, K., & Pettigrew, T. (1954). The nature of prejudice.
- Alter, A. L., & Oppenheimer, D. M. (2006). Predicting short-term stock fluctuations by using processing fluency. *Proceedings of the National Academy of Sciences*, 103(24), 9369–9372. <https://doi.org/10.1073/pnas.0601071103>
- Atwood, S., & Axt, J. R. (2021). Assessing implicit attitudes about androgyny. *Journal of Experimental Social Psychology*, 96, 104162. <https://doi.org/10.1016/j.jesp.2021.104162>
- Audacity Team. (2020). Audacity (R): free audio editor and recorder [computer program]. Version 2.1. 0 (2014).
- Auster, C. J., & Mansbach, C. S. (2012). The gender marketing of toys: An analysis of color and type of toy on the Disney store website. *Sex Roles*, 67(7–8), 375–388.
<https://doi.org/10.1007/s11199-012-0177-8>
- Banaji, M. R., & Hardin, C. D. (1996). Automatic stereotyping. *Psychological Science*, 7(3), 136–141. <https://doi.org/10.1111/j.1467-9280.1996.tb00346.x>
- Bartneck, C., Nomura, T., Kanda, T., Suzuki, T., & Kennsuke, K. (2005). A cross-cultural study on attitudes towards robots. *Proceedings of the HCI International*, Las Vegas.
- Basow, S. A., & Howe, K. G. (1987). Evaluations of college professors: Effects of professors' sex-type, and sex, and students' sex. *Psychological Reports*, 60(2), 671–678.
<https://doi.org/10.2466/pr0.1987.60.2.671>

- Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. *arXiv preprint arXiv:1506.04967*.
- Beasley, B., & Collins Standley, T. (2002). Shirts vs. skins: Clothing as an indicator of gender role stereotyping in video games. *Mass Communication and Society*, 5(3), 279–293. https://doi.org/10.1207/s15327825mcs0503_3
- Belke, B., Leder, H., Strobach, T., & Carbon, C. C. (2010). Cognitive fluency: High-level processing dynamics in art appreciation. *Psychology of Aesthetics, Creativity, and the Arts*, 4(4), 214–222. <https://doi.org/10.1037/a0019648>
- Bem, S. L. (1974). The measurement of psychological androgyny. *Journal of Consulting and Clinical Psychology*, 42(2), 155–162. <https://doi.org/10.1037/h0036215>
- Bem, S. L. (1981a). *Bem sex role inventory: Professional Manual*. Palo Alto, CA: Consulting Psychologists Press, Inc
- Bem, S. L. (1981). Gender schema theory: A cognitive account of sex typing. *Psychological Review*, 88(4), 354–364. <https://doi.org/10.1037/0033-295x.88.4.354>
- Bem, S. L. (1993). *The Lenses of Gender: Transforming the Debate on Sexual Inequality*. New Haven, Yale University Press. doi:10.12987/9780300154252
- Billger, S. M. (2009). On reconstructing school segregation: The efficacy and equity of single-sex schooling. *Economics of Education Review*, 28(3), 393–402. <https://doi.org/10.1016/j.econedurev.2007.08.005>
- Boersma, P., & Weenink, D. (2020). Praat: Doing phonetics by computer, version 6.1. 16. *Amsterdam: Phonetic Sciences, University of Amsterdam*.
- Bohn, D. (2019, January 4). *Amazon says 100 million Alexa devices have been sold - what's next?* The Verge. Retrieved November 3, 2021, from

- <https://www.theverge.com/2019/1/4/18168565/amazon-alexa-devices-how-many-sold-number-100-million-dave-limp>.
- Boroditsky, L., Schmidt, L. A., & Phillips, W. (2003). Sex, syntax, and semantics. *Language in Mind: Advances in the Study of Language and Thought*, 61-79.
- Bort, J. (2015, April 16). *Ginni Rometty is now referring to IBM's Super Smart Computer Watson as a 'he'*. Business Insider. Retrieved November 3, 2021, from <https://www.businessinsider.com/ginni-rometty-refers-to-watson-as-a-he-2015-4>.
- Brahnam, S., & de Angeli, A. (2008). Special issue on the abuse and misuse of social agents. *Interacting with Computers*, 20(3), 287–291.
<https://doi.org/10.1016/j.intcom.2008.02.001>
- Bray, F. (2007). Gender and technology. *Annual Review of Anthropology*, 36(1), 37–53.
<https://doi.org/10.1146/annurev.anthro.36.081406.094328>
- Buhrmester, M. D., Talaifar, S., & Gosling, S. D. (2018). An evaluation of Amazon's Mechanical Turk, its rapid rise, and its effective use. *Perspectives on Psychological Science*, 13(2), 149–154. <https://doi.org/10.1177/1745691617706516>
- Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk. *Perspectives on Psychological Science*, 6(1), 3–5. <https://doi.org/10.1177/1745691610393980>
- Bussey, K., & Bandura, A. (1999). Social cognitive theory of gender development and differentiation. *Psychological Review*, 106(4), 676–713. <https://doi.org/10.1037/0033-295x.106.4.676>
- Butler, J. (2004). *Undoing gender*. Routledge.
- Carli, L. L., & Eagly, A. H. (2001). Gender, hierarchy, and leadership: An introduction. *Journal of Social Issues*, 57(4), 629–636. <https://doi.org/10.1111/0022-4537.00232>

- Carpenter, J., Davis, J. M., Erwin-Stewart, N., Lee, T. R., Bransford, J. D., & Vye, N. (2009). Gender representation and humanoid robots designed for domestic use. *International Journal of Social Robotics*, 1(3), 261–265. <https://doi.org/10.1007/s12369-009-0016-4>
- Cohen, S. L., & Bunker, K. A. (1975). Subtle effects of sex role stereotypes on recruiters' hiring decisions. *Journal of Applied Psychology*, 60(5), 566–572. <https://doi.org/10.1037/0021-9010.60.5.566>
- Collins, G. S., & Moons, K. G. (2019). Reporting of artificial intelligence prediction models. *The Lancet*, 393(10181), 1577-1579.
- Crawford, J. T., Leynes, P. A., Mayhorn, C. B., & Bink, M. L. (2004). Champagne, beer, or coffee? A corpus of gender-related and neutral words. *Behavior Research Methods, Instruments, & Computers*, 36(3), 444–458. <https://doi.org/10.3758/bf03195592>
- Dacakis, G. (2000). Long-term maintenance of fundamental frequency increases in male-to-female transsexuals. *Journal of Voice*, 14(4), 549–556. [https://doi.org/10.1016/s0892-1997\(00\)80010-7](https://doi.org/10.1016/s0892-1997(00)80010-7)
- Dasgupta, N., & Asgari, S. (2004). Seeing is believing: Exposure to counterstereotypic women leaders and its effect on the malleability of automatic gender stereotyping. *Journal of Experimental Social Psychology*, 40(5), 642–658. <https://doi.org/10.1016/j.jesp.2004.02.003>
- Dasgupta, N., & Greenwald, A. G. (2001). On the malleability of automatic attitudes: Combating automatic prejudice with images of admired and disliked individuals. *Journal of Personality and Social Psychology*, 81(5), 800–814. <https://doi.org/10.1037/0022-3514.81.5.800>

- De Angeli, A. (2009). Ethical implications of verbal disinhibition with conversational agents. *PsychNology Journal*, 7(1).
- Deaux, K., & Major, B. (1987). Putting gender into context: An interactive model of gender-related behavior. *Psychological Review*, 94(3), 369–389. <https://doi.org/10.1037/0033-295x.94.3.369>
- Diekmann, A. B., & Eagly, A. H. (2000). Stereotypes as dynamic constructs: Women and men of the past, present, and future. *Personality and Social Psychology Bulletin*, 26(10), 1171–1188. <https://doi.org/10.1177/0146167200262001>
- Diekmann, A. B., & Goodfriend, W. (2006). Rolling with the changes: A role congruity perspective on gender norms. *Psychology of Women Quarterly*, 30(4), 369–383. <https://doi.org/10.1111/j.1471-6402.2006.00312.x>
- Diekmann, A. B., & Hirnisey, L. (2007). The effect of context on the silver ceiling: A role congruity perspective on prejudiced responses. *Personality and Social Psychology Bulletin*, 33(10), 1353–1366. <https://doi.org/10.1177/0146167207303019>
- Diekmann, A., & Schmader, T. (2021). Gender as embedded social cognition. <https://doi.org/10.31234/osf.io/wvx2s>
- Dovidio, J. F., & Gaertner, S. L. (1986). *Prejudice, discrimination, and racism*. Academic Press.
- Dovidio, J. F., Gaertner, S. L., & Validzic, A. (1998). Intergroup bias: Status, differentiation, and a common in-group identity. *Journal of Personality and Social Psychology*, 75(1), 109–120. <https://doi.org/10.1037/0022-3514.75.1.109>
- Eagly, A. H. (1987). Reporting sex differences. *American Psychologist*, 42(7), 756–757. <https://doi.org/10.1037/0003-066x.42.7.755>

- Eagly, A. H., & Karau, S. J. (2002). Role congruity theory of prejudice toward female leaders. *Psychological Review*, 109(3), 573–598. <https://doi.org/10.1037/0033-295x.109.3.573>
- Eagly, A. H., & Mladinic, A. (1989). Gender stereotypes and attitudes toward women and men. *Personality and Social Psychology Bulletin*, 15(4), 543–558. <https://doi.org/10.1177/0146167289154008>
- Eagly, A. H., & Mladinic, A. (1994). Are people prejudiced against women? Some answers from research on attitudes, gender stereotypes, and judgments of competence. *European Review of Social Psychology*, 5(1), 1–35. <https://doi.org/10.1080/14792779543000002>
- Eagly, A. H., Nater, C., Miller, D. I., Kaufmann, M., & Sczesny, S. (2020). Gender stereotypes have changed: A cross-temporal meta-analysis of U.S. public opinion polls from 1946 to 2018. *American Psychologist*, 75(3), 301–315. <https://doi.org/10.1037/amp0000494>
- Eagly, A. H., & Steffen, V. J. (1984). Gender stereotypes stem from the distribution of women and men into social roles. *Journal of Personality and Social Psychology*, 46(4), 735–754. <https://doi.org/10.1037/0022-3514.46.4.735>
- Eagly, A. H., & Wood, W. (1999). The origins of sex differences in human behavior: Evolved dispositions versus social roles. *American Psychologist*, 54(6), 408–423. <https://doi.org/10.1037/0003-066x.54.6.408>
- Eagly, A. H., & Wood, W. (2011). Social role theory. *Handbook of theories in social psychology*, 2, 458–476.
- Eagly, A. H., & Wood, W. (2016). Social role theory of sex differences. *The Wiley Blackwell encyclopedia of gender and sexuality studies*, 1–3.

- Eyssel, F., & Hegel, F. (2012). (S)he's got the look: Gender stereotyping of robots ¹. *Journal of Applied Social Psychology*, 42(9), 2213–2230. <https://doi.org/10.1111/j.1559-1816.2012.00937.x>
- Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, 41(4), 1149–1160. <https://doi.org/10.3758/brm.41.4.1149>
- Fessler, L. (2017). *We tested bots like Siri and Alexa to see who would stand up to sexual harassment*. Quartz. Retrieved November 3, 2021, from <https://qz.com/911681/we-tested-apples-siri-amazon-echos-alexa-microsofts-cortana-and-googles-google-home-to-see-which-personal-assistant-bots-stand-up-for-themselves-in-the-face-of-sexual-harassment/>.
- Fiske, S. T., & Stevens, L. E. (1993). *What's so special about sex? Gender stereotyping and discrimination*. Sage Publications, Inc.
- Fiske, S. T., Cuddy, A. J. C., 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), 878–902. <https://doi.org/10.1037/0022-3514.82.6.878>
- Fiske, S. T., Cuddy, A. J., & Glick, P. (2007). Universal dimensions of social cognition: warmth and competence. *Trends in Cognitive Sciences*, 11(2), 77–83. <https://doi.org/10.1016/j.tics.2006.11.005>
- Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2012). *Applied longitudinal analysis* (Vol. 998). John Wiley & Sons.

- Gaertner, S. L., Dovidio, J. F., Banker, B. S., Houlette, M., Johnson, K. M., & McGlynn, E. A. (2000). Reducing intergroup conflict: From superordinate goals to decategorization, recategorization, and mutual differentiation. *Group Dynamics: Theory, Research, and Practice*, 4(1), 98–114. <https://doi.org/10.1037/1089-2699.4.1.98>
- Galinsky, A. D., Hall, E. V., & Cuddy, A. J. C. (2013). Gendered races: Implications for interracial marriage, leadership selection, and athletic participation. *Psychological Science*, 24(4), 498–506. <https://doi.org/10.1177/0956797612457783>
- Gepperth, A., & Karaoguz, C. (2016). A bio-inspired incremental learning architecture for applied perceptual problems. *Cognitive Computation*, 8(5), 924-934.
- Glick, P., & Fiske, S. T. (2001). An ambivalent alliance: Hostile and benevolent sexism as complementary justifications for gender inequality. *American Psychologist*, 56(2), 109–118. <https://doi.org/10.1037/0003-066x.56.2.109>
- Glick, P., Zion, C., & Nelson, C. (1988). What mediates sex discrimination in hiring decisions? *Journal of Personality and Social Psychology*, 55(2), 178–186. <https://doi.org/10.1037/0022-3514.55.2.178>
- Goff, P. A., Thomas, M. A., & Jackson, M. C. (2008). “Ain’t I a Woman?”: Towards an intersectional approach to person perception and group-based harms. *Sex Roles*, 59(5–6), 392–403. <https://doi.org/10.1007/s11199-008-9505-4>
- Haraway, D. (1985). A Manifesto for Cyborgs: Science.
- Haraway, D. (1987). A manifesto for cyborgs: Science, technology, and socialist feminism in the 1980s. *Australian Feminist Studies*, 2(4), 1-42.
- Haraway, D. J. (1997). The virtual speculum in the new world order. *Feminist Review*, 55(1), 22-72.

- Heilman, M. E., Wallen, A. S., Fuchs, D., & Tamkins, M. M. (2004). Penalties for success: reactions to women who succeed at male gender-typed tasks. *Journal of Applied Psychology, 89*(3), 416–427. <https://doi.org/10.1037/0021-9010.89.3.416>
- Hester, N., Jones, B. C., & Hehman, E. (2021). Perceived femininity and masculinity contribute independently to facial impressions. *Journal of Experimental Psychology: General, 150*(6), 1147–1164. <https://doi.org/10.1037/xge0000989>
- Hoffman, C., & Hurst, N. (1990). Gender stereotypes: Perception or rationalization? *Journal of Personality and Social Psychology, 58*(2), 197–208. <https://doi.org/10.1037/0022-3514.58.2.197>
- Hoffmann, A., & Musch, J. (2018). Prejudice against women leaders: Insights from an indirect questioning approach. *Sex Roles, 80*(11–12), 681–692. <https://doi.org/10.1007/s11199-018-0969-6>
- Jackson, L. A. (1983). The perception of androgyny and physical attractiveness. *Personality and Social Psychology Bulletin, 9*(3), 405–413. <https://doi.org/10.1177/0146167283093011>
- Jones, T. (2015). Toward a description of African American vernacular English dialect regions using “Black Twitter.” *American Speech, 90*(4), 403–440. <https://doi.org/10.1215/00031283-3442117>
- Jost, J. T. (2006). The end of the end of ideology. *American Psychologist, 61*(7), 651–670. <https://doi.org/10.1037/0003-066x.61.7.651>
- Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology, 103*(1), 54–69. <https://doi.org/10.1037/a0028347>

- Jung, K., Shavitt, S., Viswanathan, M., & Hilbe, J. M. (2014). Female hurricanes are deadlier than male hurricanes. *Proceedings of the National Academy of Sciences*, 111(24), 8782–8787. <https://doi.org/10.1073/pnas.1402786111>
- Junger, J., Pauly, K., Bröhr, S., Birkholz, P., Neuschaefer-Rube, C., Kohler, C., Schneider, F., Derntl, B., & Habel, U. (2013). Sex matters: Neural correlates of voice gender perception. *NeuroImage*, 79, 275–287. <https://doi.org/10.1016/j.neuroimage.2013.04.105>
- Keijsers, M., & Bartneck, C. (2018, February). Mindless robots get bullied. In *Proceedings of the 2018 acm/ieee international conference on human-robot interaction* (pp. 205–214).
- Kite, M. E., & Deaux, K. (1987). Gender belief systems: Homosexuality and the implicit inversion theory. *Psychology of Women Quarterly*, 11(1), 83–96. <https://doi.org/10.1111/j.1471-6402.1987.tb00776.x>
- Koch, A. J., D’Mello, S. D., & Sackett, P. R. (2015). A meta-analysis of gender stereotypes and bias in experimental simulations of employment decision making. *Journal of Applied Psychology*, 100(1), 128–161. <https://doi.org/10.1037/a0036734>
- Koenig, A. M., Eagly, A. H., Mitchell, A. A., & Ristikari, T. (2011). Are leader stereotypes masculine? A meta-analysis of three research paradigms. *Psychological Bulletin*, 137(4), 616–642. <https://doi.org/10.1037/a0023557>
- Kurzban, R., Tooby, J., & Cosmides, L. (2001). Can race be erased? Coalitional computation and social categorization. *Proceedings of the National Academy of Sciences*, 98(26), 15387–15392. <https://doi.org/10.1073/pnas.251541498>
- Lai, C. K., Marini, M., Lehr, S. A., Cerruti, C., Shin, J. E. L., Joy-Gaba, J. A., Ho, A. K., Teachman, B. A., Wojcik, S. P., Koleva, S. P., Frazier, R. S., Heiphetz, L., Chen, E. E., Turner, R. N., Haidt, J., Kesebir, S., Hawkins, C. B., Schaefer, H. S., Rubichi, S., . . .

- Nosek, B. A. (2014). Reducing implicit racial preferences: I. A comparative investigation of 17 interventions. *Journal of Experimental Psychology: General*, 143(4), 1765–1785.
<https://doi.org/10.1037/a0036260>
- Legg, S., & Hutter, M. (2007). A collection of definitions of intelligence. *Frontiers in Artificial Intelligence and applications*, 157, 17.
- Levon, E. (2014). Categories, stereotypes, and the linguistic perception of sexuality. *Language in Society*, 43(5), 539–566. <https://doi.org/10.1017/s0047404514000554>
- Levon, E. (2015). Integrating intersectionality in language, gender, and sexuality research. *Language and Linguistics Compass*, 9(7), 295–308. <https://doi.org/10.1111/lnc3.12147>
- Lick, D. J., & Johnson, K. L. (2013). Fluency of visual processing explains prejudiced evaluations following categorization of concealable identities. *Journal of Experimental Social Psychology*, 49(3), 419–425. <https://doi.org/10.1016/j.jesp.2013.01.001>
- Lick, D. J., & Johnson, K. L. (2015). The interpersonal consequences of processing ease. *Current Directions in Psychological Science*, 24(2), 143–148.
<https://doi.org/10.1177/0963721414558116>
- Lick, D. J., Johnson, K. L., & Rule, N. O. (2015). Disfluent processing of nonverbal cues helps to explain anti-bisexual prejudice. *Journal of Nonverbal Behavior*, 39(3), 275–288.
<https://doi.org/10.1007/s10919-015-0211-y>
- Linville, S. E. (1998). Acoustic correlates of perceived versus actual sexual orientation in men's Speech. *Folia Phoniatica et Logopaedica*, 50(1), 35–48.
<https://doi.org/10.1159/000021447>

- Lippa, R., & Connelly, S. (1990). Gender diagnosticity: A new Bayesian approach to gender-related individual differences. *Journal of Personality and Social Psychology*, 59(5), 1051–1065. <https://doi.org/10.1037/0022-3514.59.5.1051>
- Litman, L., Robinson, J., & Abberbock, T. (2017). TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods*, 49(2), 433–442. <https://doi.org/10.3758/s13428-016-0727-z>
- Maccoby, E. E., & Jacklin, C. N. (1987). Gender segregation in childhood. *Advances in Child Development and Behavior*, 20, 239-287.
- Mack, S., & Munson, B. (2012). The influence of /s/ quality on ratings of men’s sexual orientation: Explicit and implicit measures of the ‘gay lisp’ stereotype. *Journal of Phonetics*, 40(1), 198–212. <https://doi.org/10.1016/j.wocn.2011.10.002>
- Macrae, C. N., & Bodenhausen, G. V. (2000). Social cognition: Thinking categorically about others. *Annual Review of Psychology*, 51(1), 93–120. <https://doi.org/10.1146/annurev.psych.51.1.93>
- Mahalik, J. R. (2000). Gender role conflict in men as a predictor of self-ratings of behavior on the interpersonal circle. *Journal of Social and Clinical Psychology*, 19(2), 276–292. <https://doi.org/10.1521/jscp.2000.19.2.276>
- Major, B., Carnevale, P. J. D., & Deaux, K. (1981). A different perspective on androgyny: Evaluations of masculine and feminine personality characteristics. *Journal of Personality and Social Psychology*, 41(5), 988–1001. <https://doi.org/10.1037/0022-3514.41.5.988>
- Martin, A. E., & Mason, M. F. (2022). What does it mean to be (seen as) human? The importance of gender in humanization. *Journal of Personality and Social Psychology*. <https://doi.org/10.1037/pspa0000293>

- Martin, A. E., & Slepian, M. L. (2018). Dehumanizing gender: The debiasing effects of gendering human-abstracted entities. *Personality and Social Psychology Bulletin*, 44(12), 1681–1696. <https://doi.org/10.1177/0146167218774777>
- Martin, A. E., & Slepian, M. L. (2020). The primacy of gender: Gendered cognition underlies the big two dimensions of social cognition. *Perspectives on Psychological Science*, 16(6), 1143–1158. <https://doi.org/10.1177/1745691620904961>
- Martin, C. L., & Halverson, C. F. (1983). The effects of sex-typing schemas on young children's memory. *Child Development*, 54(3), 563. <https://doi.org/10.2307/1130043>
- McCreary, D. R. (1994). The male role and avoiding femininity. *Sex Roles*, 31(9-10), 517-531.
- Misersky, J., Gyga, P. M., Canal, P., Gabriel, U., Garnham, A., Braun, F., Chiarini, T., Englund, K., Hanulíková, A., Öttl, A., Valdrova, J., von Stockhausen, L., & Sczesny, S. (2013). Norms on the gender perception of role nouns in Czech, English, French, German, Italian, Norwegian, and Slovak. *Behavior Research Methods*. <https://doi.org/10.3758/s13428-013-0409-z>
- Morgenroth, T., & Ryan, M. K. (2018). Gender trouble in social psychology: How can Butler's work inform experimental social psychologists' conceptualization of gender? *Frontiers in Psychology*, 9. <https://doi.org/10.3389/fpsyg.2018.01320>
- Morgenroth, T., & Ryan, M. K. (2020). The effects of gender trouble: An integrative theoretical framework of the perpetuation and disruption of the gender/sex binary. *Perspectives on Psychological Science*, 16(6), 1113–1142. <https://doi.org/10.1177/1745691620902442>
- Mori, M. (1970). Bukimi no tani [the uncanny valley]. *Energy*, 7, 33-35.
- Muthén, B., & Muthén, L. (2017). *Mplus*. Chapman and Hall/CRC.

- Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, 56(1), 81-103.
- 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. <https://doi.org/10.1111/j.1559-1816.1997.tb00275.x>
- Nilsson, N. J. (1982). *Principles of artificial intelligence*. Springer Science & Business Media.
- Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002). Math = male, me = female, therefore math $\neq$ me. *Journal of Personality and Social Psychology*, 83(1), 44–59. <https://doi.org/10.1037/0022-3514.83.1.44>
- Nosek, B. A., Smyth, F. L., Hansen, J. J., Devos, T., Lindner, N. M., Ranganath, K. A., Smith, C. T., Olson, K. R., Chugh, D., Greenwald, A. G., & Banaji, M. R. (2007). Pervasiveness and correlates of implicit attitudes and stereotypes. *European Review of Social Psychology*, 18(1), 36–88. <https://doi.org/10.1080/10463280701489053>
- Parisi, G. I., Kemker, R., Part, J. L., Kanan, C., & Wermter, S. (2019). Continual lifelong learning with neural networks: A review. *Neural Networks*, 113, 54-71.
- Perez, S. (2017, July 11). *Siri usage and engagement dropped since last year, as Alexa and Cortana grew*. TechCrunch. Retrieved November 3, 2021, from <https://techcrunch.com/2017/07/11/siri-usage-and-engagement-dropped-since-last-year-as-alexa-and-cortana-grew/>.
- Pettigrew, T. F., & Tropp, L. R. (2008). How does intergroup contact reduce prejudice? Meta-analytic tests of three mediators. *European Journal of Social Psychology*, 38(6), 922–934. <https://doi.org/10.1002/ejsp.504>

- Plant, E. A., & Devine, P. G. (1998). Internal and external motivation to respond without prejudice. *Journal of Personality and Social Psychology*, 75(3), 811–832.
<https://doi.org/10.1037/0022-3514.75.3.811>
- Poole, D. L., & Mackworth, A. K. (2010). *Artificial Intelligence: foundations of computational agents*. Cambridge University Press.
- Prentice, D. A., & Carranza, E. (2002). What women and men should be, shouldn't be, are allowed to be, and don't have to be: The contents of prescriptive gender stereotypes. *Psychology of Women Quarterly*, 26(4), 269–281. <https://doi.org/10.1111/1471-6402.t01-1-00066>
- Purdie-Vaughns, V., & Eibach, R. P. (2008). Intersectional invisibility: The distinctive advantages and disadvantages of multiple subordinate-group identities. *Sex Roles*, 59(5–6), 377–391. <https://doi.org/10.1007/s11199-008-9424-4>
- Purnell, T., Idsardi, W., & Baugh, J. (1999). Perceptual and phonetic experiments on American English dialect identification. *Journal of Language and Social Psychology*, 18(1), 10–30.
<https://doi.org/10.1177/0261927x99018001002>
- Purtill, C. (2021, February 12). *Hey, Alexa, are you sexist?* The New York Times. Retrieved November 3, 2021, from <https://www.nytimes.com/2021/02/12/us/alexa-bots-female-voice.html>.
- Remedios, J. D., Chasteen, A. L., & Paek, J. D. (2011). Not all prejudices are experienced equally: Comparing experiences of racism and sexism in female minorities. *Group Processes & Intergroup Relations*, 15(2), 273–287.
<https://doi.org/10.1177/1368430211411594>

- Remedios, J. D., Chasteen, A. L., Rule, N. O., & Plaks, J. E. (2011). Impressions at the intersection of ambiguous and obvious social categories: Does gay + Black = likable? *Journal of Experimental Social Psychology*, 47(6), 1312–1315.
<https://doi.org/10.1016/j.jesp.2011.05.015>
- Remedios, J. D., & Snyder, S. H. (2015). Where do we go from here? Toward an inclusive and intersectional literature of multiple stigmatization. *Sex Roles*, 73(9–10), 408–413.
<https://doi.org/10.1007/s11199-015-0543-4>
- Remez, R. E., Fellowes, J. M., & Rubin, P. E. (1997). Talker identification based on phonetic information. *Journal of Experimental Psychology: Human Perception and Performance*, 23(3), 651–666. <https://doi.org/10.1037/0096-1523.23.3.651>
- Rieger, G., Linsenmeier, J. A. W., Gygax, L., Garcia, S., & Bailey, J. M. (2008). Dissecting “gaydar”: Accuracy and the role of masculinity–femininity. *Archives of Sexual Behavior*, 39(1), 124–140. <https://doi.org/10.1007/s10508-008-9405-2>
- Robertson, J. (2010). Gendering humanoid robots: Robo-sexism in Japan. *Body & Society*, 16(2), 1–36. <https://doi.org/10.1177/1357034x10364767>
- 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), 629–645. <https://doi.org/10.1037/0022-3514.74.3.629>
- 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), 157–176. <https://doi.org/10.1037/0022-3514.87.2.157>
- Rudman, L. A., & Glick, P. (1999). Feminized management and backlash toward agentic women: The hidden costs to women of a kinder, gentler image of middle managers.

- Journal of Personality and Social Psychology*, 77(5), 1004–1010.
<https://doi.org/10.1037/0022-3514.77.5.1004>
- Rudman, L. A., & Glick, P. (2001). Prescriptive gender stereotypes and backlash toward agentic women. *Journal of Social Issues*, 57(4), 743–762. <https://doi.org/10.1111/0022-4537.00239>
- 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.
<https://doi.org/10.1016/j.jesp.2011.10.008>
- Rudman, L. A., & Phelan, J. E. (2008). Backlash effects for disconfirming gender stereotypes in organizations. *Research in Organizational Behavior*, 28, 61–79.
<https://doi.org/10.1016/j.riob.2008.04.003>
- Rudman, L. A., & Phelan, J. E. (2010). The effect of priming gender roles on women’s implicit gender beliefs and career aspirations. *Social Psychology*, 41(3), 192–202.
<https://doi.org/10.1027/1864-9335/a000027>
- Russell, S. J., & Norvig, P. (2016). *Artificial intelligence: a modern approach*. Prentice Hall Series in Artificial Intelligence.
- Sánchez Ramos, A. C., Contreras, V., Santos, A., Aguillon, C., Garcia, N., Rodriguez, J. D., ... & Strait, M. K. (2018). A preliminary study of the effects of racialization and humanness on the verbal abuse of female-gendered robots. In *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 227-228).

- Sanchez, D. T., & Bonam, C. M. (2009). To disclose or not to disclose biracial identity: The effect of biracial disclosure on perceiver evaluations and target responses. *Journal of Social Issues*, 65(1), 129–149. <https://doi.org/10.1111/j.1540-4560.2008.01591.x>
- Schaller, M., & Crandall, C. S. (Eds.). (2003). *The psychological foundations of culture*. Psychology Press.
- Scott, G. G., Keitel, A., Becirspahic, M., Yao, B., & Sereno, S. C. (2018). The Glasgow Norms: Ratings of 5,500 words on nine scales. *Behavior Research Methods*, 51(3), 1258–1270. <https://doi.org/10.3758/s13428-018-1099-3>
- Simpson, A. P. (2009). Phonetic differences between male and female speech. *Language and Linguistics Compass*, 3(2), 621–640. <https://doi.org/10.1111/j.1749-818x.2009.00125.x>
- Sirin, S., McCreary, D., & Mahalik, J. (2004). Differential reactions to men and women's gender role transgressions: Perceptions of social status, sexual orientation, and value dissimilarity. *The Journal of Men's Studies*, 12(2), 119–132. <https://doi.org/10.3149/jms.1202.119>
- Spence, J. T., & Helmreich, R. L. (1979). *Masculinity and Femininity: Their Psychological Dimensions, Correlates, and Antecedents*. University of Texas Press.
- Spence, J. T. (1993). Gender-related traits and gender ideology: Evidence for a multifactorial theory. *Journal of Personality and Social Psychology*, 64(4), 624.
- Stern, C. (2022). Political ideology and social categorization. *Advances in Experimental Social Psychology*, 64, 167-233.
- Stern, C., & Rule, N. O. (2018). Physical androgyny and categorization difficulty shape political conservatives' attitudes toward transgender people. *Social Psychological and Personality Science*, 9(1), 24–31. <https://doi.org/10.1177/1948550617703172>

- Stroessner, S. J., Benitez, J., Perez, M. A., Wyman, A. B., Carpinella, C. M., & Johnson, K. L. (2020). What's in a shape? Evidence of gender category associations with basic forms. *Journal of Experimental Social Psychology*, 87, 103915. <https://doi.org/10.1016/j.jesp.2019.103915>
- Taylor, S. E., Crocker, J., Fiske, S. T., Sprinzen, M., & Winkler, J. D. (1979). The generalizability of salience effects. *Journal of Personality and Social Psychology*, 37(3), 357. <https://doi.org/10.1037/0022-3514.37.3.357>
- Tracy, E. C., Bainter, S. A., & Satariano, N. P. (2015). Judgments of self-identified gay and heterosexual male speakers: Which phonemes are most salient in determining sexual orientation? *Journal of Phonetics*, 52, 13–25. <https://doi.org/10.1016/j.wocn.2015.04.001>
- Upbin, B. (2013, October 18). *IBM's Watson now tackles clinical trials at MD Anderson Cancer Center*. Forbes. Retrieved November 3, 2021, from <https://www.forbes.com/sites/bruceupbin/2013/10/18/ibms-watson-now-tackles-clinical-trials-at-md-anderson-cancer-center/?sh=785f38994e1e>.
- Vandello, J. A., & Bosson, J. K. (2013). Hard won and easily lost: A review and synthesis of theory and research on precarious manhood. *Psychology of Men & Masculinity*, 14(2), 101–113. <https://doi.org/10.1037/a0029826>
- Vandello, J. A., Bosson, J. K., Cohen, D., Burnaford, R. M., & Weaver, J. R. (2008). Precarious manhood. *Journal of Personality and Social Psychology*, 95(6), 1325–1339. <https://doi.org/10.1037/a0012453>
- Webster, A. (2019, January 30). *The AI video game character that influenced an industry*. The Verge. Retrieved November 3, 2021, from

<https://www.theverge.com/2019/1/30/18201130/halo-cortana-ai-assistant-siri-alexa-video-game>.

Weisgram, E. S., Fulcher, M., & Dinella, L. M. (2014). Pink gives girls permission: Exploring the roles of explicit gender labels and gender-typed colors on preschool children's toy preferences. *Journal of Applied Developmental Psychology*, 35(5), 401–409.

<https://doi.org/10.1016/j.appdev.2014.06.004>

West, M., Kraut, R., & Ei Chew, H. (2019). I'd blush if I could: closing gender divides in digital skills through education. UNESCO.

<https://unesdoc.unesco.org/ark:/48223/pf0000367416.page=1>

Westfall, J., Kenny, D. A., & Judd, C. M. (2014). Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology: General*, 143(5), 2020–2045.

<https://doi.org/10.1037/xge0000014>

White, M. J., & White, G. B. (2006). Implicit and explicit occupational gender stereotypes. *Sex Roles*, 55(3–4), 259–266. <https://doi.org/10.1007/s11199-006-9078-z>

Wilkie, J. E. B., & Bodenhausen, G. V. (2012). Are numbers gendered? *Journal of Experimental Psychology: General*, 141(2), 206–210. <https://doi.org/10.1037/a0024875>

Wilkie, J. E. B., & Bodenhausen, G. V. (2015). The numerology of gender: gendered perceptions of even and odd numbers. *Frontiers in Psychology*, 6.

<https://doi.org/10.3389/fpsyg.2015.00810>

Wilson, B. D. M., & Meyer, I. H. (2021). Nonbinary LGBTQ adults in the United States.

Retrieved from: <https://williamsinstitute.law.ucla.edu/publications/nonbinary-lgbtq-adults-us/>

- Wilson, J. P., Remedios, J. D., & Rule, N. O. (2017). Interactive effects of obvious and ambiguous social categories on perceptions of leadership: When double-minority status may be beneficial. *Personality and Social Psychology Bulletin*, 43(6), 888–900.
<https://doi.org/10.1177/0146167217702373>
- Winkielman, P., Halberstadt, J., Fazendeiro, T., & Catty, S. (2006). Prototypes are attractive because they are easy on the mind. *Psychological Science*, 17(9), 799–806.
<https://doi.org/10.1111/j.1467-9280.2006.01785.x>
- Wood, W., & Eagly, A. H. (2015). Two traditions of research on gender identity. *Sex Roles*, 73(11–12), 461–473. <https://doi.org/10.1007/s11199-015-0480-2>