Received: 16 September 2024 | Accepted: 29 September 2024

DOI: 10.1002/arcp.1102

REVIEW

Reducing prejudice with counter-stereotypical AI

Erik Hermann[1] | Julian De Freitas[2] | Stefano Puntoni[3]

[1]ESCP Business School, Berlin, Germany

[2]Harvard Business School, Harvard University, Boston, Massachusetts, USA

[3]Wharton School of Business, University of Pennsylvania, Philadelphia, Pennsylvania, USA

Correspondence
Erik Hermann, ESCP Business School, Heubnerweg 8-10, 14093 Berlin, Germany.
Email: ehermann@escp.eu

Abstract

Based on a review of relevant literature, we propose that the proliferation of AI with human-like and social features presents an unprecedented opportunity to address the underlying cognitive and affective drivers of prejudice. An approach informed by the psychology of intergroup contact and prejudice reduction is necessary because current AI systems often reinforce or avoid prejudices. Against this backdrop, we outline unique opportunities for prejudice reduction through 'synthetic' intergroup contact, wherein consumers interact with AI products and services that counter stereotypes and serve as a 'proxy' members of the outgroup (i.e., counter-stereotypical AI). In contrast to human-human contact, humanizing and socializing AI can reduce prejudice through more repeated, direct, unavoidable, private, non-judgmental, collaborative, and need-satisfying contact. We illustrate the potential of synthetic intergroup contact with counter-stereotypical AI using examples of gender stereotypes and hate speech and discuss practical considerations for implementing counter-stereotypical AI without inadvertently perpetuating or reinforcing prejudice.

KEYWORDS

artificial intelligence, intergroup psychology, prejudice, stereotypes, synthetic contact

1 | INTRODUCTION

There have been various calls in both the private and public sectors to improve the representation of diverse groups, and principles of diversity, equity, and inclusion have become hot-button topics (Bernstein et al., 2020; Kipnis et al., 2021; Park et al., 2023). Yet, the fact remains that on many key metrics (e.g., the share of minorities in management) society has seen little progress (Kraus et al., 2022; Torrez et al., 2024). One plausible hypothesis is that many of the espoused interventions that tell people what biases they have and must fix, as in so-called "diversity training" or "anti-bias training", simply do not work on their own and often backfire, for instance, by making attendees more defensive and reinforcing any biases they might have (Dobbin & Kalev, 2018; Paluck et al., 2021). Even when such training has a positive effect, these effects are often small, short-lasting, and do not translate into organizational change (Chang et al., 2019; Dobbin & Kalev, 2018; Paluck et al., 2021). While such training is premised on the assumption that one can reduce prejudice through an hour, week, or month of short-term educational interventions, the reality is that prejudice is caused by stereotypes that become ingrained through years of cultural and media exposure to stereotyped portrayals of social groups. Reducing prejudice therefore requires revising stereotypes and affective reactions to outgroups through prolonged, substantial exposure to counter-stereotypical instances.

The proliferation of AI with human-like features within the fabric of daily life provides an unprecedented opportunity to satisfy these requirements and revise the underlying cognitive and affective drivers of social prejudice, because of AI's unique features. However, realizing this potential requires an approach informed by the psychology of intergroup contact and prejudice reduction. To wit, current AI devices, applications, and other systems tend to either 'reinforce' or 'avoid' prejudices. For instance, intelligent personal assistants like Amazon's Alexa or Apple's Siri often act as subservient synthetic females receiving and obeying orders (Martin & Mason, 2023), thereby 'reinforcing' gender-stereotypical social subordination (Fortunati et al., 2022). As another example, generative AI platforms regularly deal with hate speech directed against members of different social groups with strict censorship (Cobbe, 2021; Hangartner et al., 2021), thereby 'avoiding' rather than addressing the prejudices that underlie hate speech.

Against this backdrop, we review relevant literature to lay out unique opportunities for prejudice reduction through 'synthetic intergroup contact'—in which consumers' interactions with an AI that counters stereotypes lead to stereotype reduction by serving as a 'proxy' member of the outgroup—and we focus on those areas where AI is likely to provide more potent effects than the status quo of contact with human members of stereotyped groups. First, we illustrate consumers' differential responses to AI and the phenomena of 'humanizing' and 'socializing' AI products and service. Second, we review how these phenomena enable these systems to serve as counter-stereotypes for reducing prejudice and provide concrete examples of gendered AI and hate speech moderation. Third, we propose that AI can be a potent means of 'synthetic' intergroup contact, given that consumers' contact with AI can be more (i) repeated, (ii) direct, (iii) unavoidable, (iv) private, (v) non-judgmental, (vi) collaborative, and (vii) need-satisfying than contact with human outgroup members. Fourth, and finally, we point out how to implement counter-stereotypical AI to accomplish intended societal effects without inadvertently perpetuating prejudices.

2 | AI, CONSUMERS, AND SYNTHETIC 'INTERGROUP' CONTACT

AI has become ubiquitous in our lives. Interactions with AI can be regular, frequent, diverse, intensive, and relational in nature (De Freitas et al., 2024; Martin & Mason, 2023; Uysal et al., 2022) and occur in private, social, and professional contexts, wherein AI can take on various instantiations. These include intelligent personal assistants (Amazon's Alexa, Apple's Siri), service chatbots (Bank of America's Erica, Lemonade's Maya), personal productivity tools (OpenAI's ChatGPT, Google's Gemini), smart products (e.g., fitness trackers, smart home devices), companion AIs (Replika, Anima), AI tutors (Georgia Tech's AskJill, Kahn Academy's Kahnmigo), and more (e.g., financial robo advisors).

For the most part, consumer research on AI offerings has focused on negative and positive psychological responses to the technology. Prior research has shown that consumers are often reluctant to use AI or reject its advice altogether (i.e., "algorithm aversion;"; De Freitas et al., 2023; Dietvorst et al., 2015). The reasons for this aversion are manifold and include but are not limited to consumers' perception of AI as less authentic and moral (Bigman & Gray, 2018; Dietvorst & Bartels, 2022; Giroux et al., 2022; Jago, 2019; Jago et al., 2022), neglecting their unique needs (Longoni et al., 2019), not learning from mistakes (Reich et al., 2023), being less capable than humans at the same tasks (Agarwal et al., 2024), and is viewed as not a member of the same species as humans (De Freitas et al., 2023).

Conversely, consumers tend to react more positively to AI in certain circumstances, such as when they are provided control and oversight (Dietvorst et al., 2018; Longoni & Cian, 2022) or with explanations of AI functioning (Berger et al., 2021; Cadario et al., 2021; Chen et al., 2023; Clegg et al., 2023), and when it is anthropomorphized (Blut et al., 2021; Castelo et al., 2019) or socially present (i.e., consumers perceive a sense of being with another being; Flavián et al., 2024). Consumers also respond more favorably toward AI when the task is objective (vs. subjective; Castelo et al., 2019), when their needs are certain (Zhu et al., 2022), when they feel less judged than by humans (Duani et al., 2024), and when offers are worse than expected (Garvey et al., 2023). Besides, more positive consumer responses have been found in embarrassing service encounters (Holthöwer & van Doorn, 2022; Pitardi et al., 2022), in utilitarian (vs. hedonic) contexts (Longoni & Cian, 2022; Wien & Peluso, 2021), and for search (vs. experience) products (Xie et al., 2022).

Yet, for our purposes, two trends are notable from this prior work. First, AI products and services are 'humanized,' that is, equipped with human characteristics like voice, names, physical appearances, or simulated emotions and mental states (Blut et al., 2021). These humanlike features turn human-technology interactions into human-human-like interactions, elicit social responses from users, and shape intergroup relations and attitudes (Jackson et al., 2020; Nass et al., 1994; Nass & Moon, 2000). For instance, interactions with AI-based conversational interfaces like chatbots, intelligent personal assistants, or robo-advisors can mimic the form of human-to-human conversations (Bergner et al., 2023; Hildebrand & Bergner, 2021). Due to reciprocal communication and repeated, conversations with increasingly sophisticated technology, consumers can have social experiences with AI systems (Pantano & Scarpi, 2022; Puntoni et al., 2021) and even form emotional connections with them (De Freitas et al., 2024; Yu & Fan, 2024).

Second, AI is 'socialized.' That is, AI can employ markers that evoke social group categories (Davis et al., 2023; Martin & Mason, 2023; Yi & Turner, 2024). Because the tendency to stereotype is so pervasive and ubiquitous in humans, AIs that employ social features can serve as 'proxy members' of human social groups. For example, to evoke gender-stereotypical perceptions of AI, it suffices to employ simple cues such as colors (pink vs. blue), female/male voices or names, or physical appearances like long versus short hair (Ahn et al., 2022; Lee et al., 2024). These cues of AI mean that interactions with AI may also trigger intergroup social perceptions and cognitions like prejudices. In this way, AI can serve as a means of 'synthetic' intergroup contact.

3 | THE OPPORTUNITY: COUNTER-STEREOTYPICAL AI

Given the phenomenon of synthetic intergroup contact, here we focus on the opportunity for 'counter-stereotypical AI', which we define as AI design and deployment that embodies characteristics, roles, or decisions that diverge from societal stereotypes associated with certain groups—for instance, a male virtual assistant programmed to express warmth and empathy, or a member of a racial or ethnic minority designed to be a competent virtual real estate agent. A key contention of our work is that by presenting attributes, behaviors, or outcomes that counter common stereotypes, these AI systems have the potential to intervene in the psychological processes underlying stereotypical associations, thereby promoting long-lasting prejudice reduction.

Given the ubiquity and pervasiveness of stereotypical cognition (Fiske et al., 2007; Martin & Slepian, 2021), we expect that consumers are likely to readily recognize counter-stereotypical instances. Stereotype-disconfirming attributes are likely to be particularly salient when they occur in domains where an existing stereotype applies more strongly (Dasgupta & Asgari, 2004). For instance, the category of education services elicits more competence associations than does the category of care work, such that the existence of an AI math tutor with a Black personality is more likely to be perceived as counter-stereotypical. Before we shed light on the AI-unique opportunities for prejudice reduction, we briefly illustrate the psychological processes underlying prejudice reduction and how they would be implicated in the concrete examples of gendered AI and hate speech moderation.

3.1 | Prejudices and intergroup contact

Prejudices are negative biases toward a social category of people (Paluck & Green, 2009) that consist of both cognitive and affective dimensions. The cognitive dimension involves beliefs about the characteristics, attributes, and behaviors of members of certain groups that are likely to have negative connotations when they relate to outgroup members (i.e., stereotypes; Hilton & von Hippel, 1996). The affective dimension involves anticipated emotions when interacting with outgroup members as well as feelings of positivity or negativity toward them (i.e., favourability; Tropp & Pettigrew, 2005). One of the most effective approaches for prejudice reduction is intergroup contact: the actual or symbolic interaction between representatives of different social groups (Intergroup Contact Theory; Allport, 1954). Positive contact effects occur across many cultures and social groups, which can either be directly involved in the intergroup encounter or not (Pettigrew, 2009; Pettigrew & Tropp, 2006). Examples of indirect contact include extended contact (knowing that other members of your ingroup have relationships with outgroup members), imagined contact (mentally simulating positive intergroup interactions), parasocial contact (being exposed to media information and portrayals about outgroup members), and virtual contact via computer-enabled communication (Dovidio et al., 2017; Hodson et al., 2018). Intergroup contact reduces prejudice by intervening on the cognitive and affective processes underlying prejudice (Crisp & Turner, 2011; Dovidio et al., 2017; Pettigrew & Tropp, 2008).

3.2 | Cognitive processes underlying prejudice reduction

Consumers are likely to perceive counter-stereotypical AI products and service encounters as surprising, unfamiliar, and incongruent conjunction of social categories (Hutter et al., 2009; Hutter & Crisp, 2005; Kunda et al., 1990). Incongruent combinations of social categories (e.g., a highly competent black tutor, or dark-skinned immigration avatar deployed on behalf of the U.S. government) challenge stereotypical expectations (e.g., that black people cannot be educators, or that Mexican immigrants are not ideal U.S. citizens), contradicting prevailing negative stereotypes stored in memory (Crisp & Turner, 2011; Hutter & Crisp, 2005). If consumers are unable to make sense of the incongruent conjunction by retrieving stored information from memory, they may resolve the inconsistency by generating new, emergent attributes (Hutter et al., 2009; Hutter & Crisp, 2005; Kunda et al., 1990). This process is termed *recategorization*. In this way, consumers exposed to counter-stereotypical AI may begin to rely less on their initial stereotypes when thinking about and judging members of certain groups (Hutter & Crisp, 2005) and attach less significance to these initial categories (Howe & Krosnik, 2017).

In the long run, as consumers are repeatedly exposed to and interact with counter-stereotypical AI, they likewise repeatedly engage in the cognitive process of inconsistency resolution and come to spontaneously inhibit stereotype-based knowledge in favor of more structured, nuanced ways of (re-)categorizing others that run counter to these simpler stereotypes (Crisp & Turner, 2011). In the ideal case, the more nuanced, counter-stereotypical way of representing others eventually supplants the stereotypical way of doing so. For example, existing research finds that long-term exposure to women in counter-stereotypical roles like leadership positions reduces gender stereotypes (Dasgupta & Asgari, 2004; Lawson et al., 2022). Similarly, we suggest that frequent synthetic contact with counter-stereotypical AI gradually attenuates stereotypes by both weakening the strength of these stereotypes and changing their content. That is, counter-stereotypical AI can induce a cognitive shift from psychologically representing certain group members in a heuristic way based on stereotypical information stored in memory (e.g., race or gender stereotypes; Brough et al., 2024), toward representing them in a more nuanced, critical, accurate manner (Crisp & Turner, 2011; Hutter et al., 2009; Hutter & Crisp, 2005).

3.2.1 | Example: gendered AI

As a concrete example of how the cognitive route leads to stereotype reduction, consider the use of gendered intelligent personal assistants. “Alexa, play some music.” “Alexa, show my calendar.” “Alexa, set the alarm for 7 a.m.” These are examples of everyday commands that users give Amazon’s artificially intelligent (voice-based) personal assistant. Intelligent personal assistants such as Alexa, Apple’s Siri, or Microsoft’s Cortana tend to have names, voices, morphology, personae, and other features associated with the female gender, as do many other AI-driven applications, including chatbots, robot receptionists/concierges, service kiosks, and more. While female gendering can increase consumers’ attachment, usage intentions, and perceived humanness of the AI (Blut et al., 2021; Borau et al., 2021; Martin & Mason, 2023), interactions with AI-driven technology often take the form of commander-servant subordination, wherein a subservient synthetic female receives and obeys orders—the “servile helper” (Martin & Mason, 2023). Thereby,

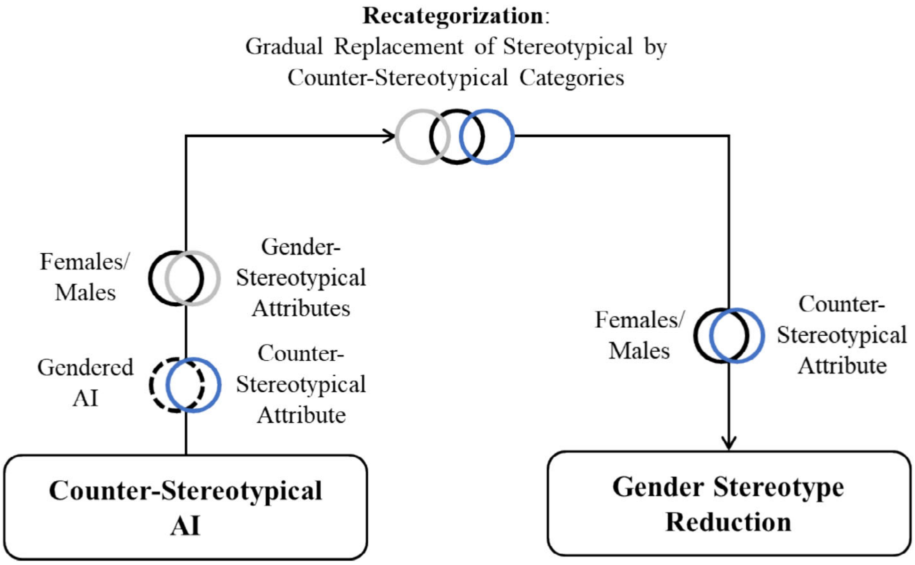
Recategorization:
Gradual Replacement of Stereotypical by Counter-Stereotypical Categories
Females/ Males
Gender-Stereotypical Attributes
Gendered AI
Counter-Stereotypical Attribute
Counter-Stereotypical AI
Females/ Males
Counter-Stereotypical Attribute
Gender Stereotype Reduction

FIGURE 1 Gender stereotype reduction through the cognitive route.

intelligent personal assistants perform stereotypical feminine roles through "digital domesticity" (Woods, 2018) and reproduce gendered dimensions of domestic labor (Schiller & McMahon, 2019) and social subordination (Fortunati et al., 2022). Intelligent personal assistants and other AI applications such as algorithms delivering gender-biased advertisements and product recommendations; Lambrecht & Tucker, 2019; Rathee et al., 2023) can thus reify and strengthen harmful gender stereotypes (Borau, 2024; Martin & Mason, 2023; Woods, 2018). In contrast, counter-stereotypical AI can help to counter gender stereotypes by triggering the cognitive recategorization process described above (see Figure 1).

Counter-stereotypical gendered AI is implemented by combining the AI's gender (i.e., female vs. male) with the opposite-gender attribute (i.e., competence vs. warmth): for instance, a female navigation device that gives assertive commands to users, or a warm male customer greeting robot. Consumers are likely to perceive such counter-stereotypical AI as incongruent conjunctions of social categories that challenge stereotypical expectations and conflict with stereotyped knowledge stored in memory. When consumers repeatedly interact with and engage in the recategorization resolution process, stereotype-based categories are gradually replaced by counter-stereotypical representations.

3.3 | Affective processes underlying prejudice reduction

Synthetic contact can also reduce prejudice through the affective subprocesses of increasing empathy (Batson et al., 1997; Stephan & Finlay, 1999; Vanman, 2016) and reducing intergroup anxiety (Stephan, 2014; Turner et al., 2007; Wölfer et al., 2019). Empathy is a form of positive affect and refers to the ability to adopt others' points of view and experience affective reactions to their observed experiences (Davis, 1994). Intergroup anxiety is a form of negative affect and relates to anticipated negative psychological, behavioral, or evaluative consequences for the self-arising from intergroup encounters (Stephan & Stephan, 1985).

Synthetic contact with counter-stereotypical AI can elicit these same affective subprocesses, increasing positive affect (i.e., empathy) and decreasing negative affect (i.e., intergroup anxiety), which in turn positively influences emotions and favorability toward outgroup members. Emotional and empathetic AI—that is, AI that is able to recognize and understand customers' emotions, provide emotion management recommendations, and establish, emotional connections with customers (Huang & Rust, 2024)—is increasingly integrated into consumption and service contexts and thus into consumer interactions with AI (Esmaeilzadeh & Vaezi, 2022; Liu-Thompkins et al., 2022). Such AIs lead consumers to respond with more effect and empathy (Nielsen et al., 2022; Pataranutaporn et al., 2023; Yang & Xie, 2024). Furthermore, research on human-human interactions finds that simply observing other humans' empathetic responses increases observers' own empathy through observational learning (Zhou et al., 2024), suggesting a similar effect could arise when observing empathic AI. Increased empathy could in turn evoke positive emotions and favorability toward the social groups with which the AI is associated, thereby reducing prejudices toward these groups.

Repeated, direct synthetic contact with counter-stereotypical AI could also mitigate consumers' anxiety about negative interactions with outgroup members, as they become increasingly familiar and comfortable with the outgroup via their interactions with the AI in situations that do not present significant social risks. Since AI is generally viewed as being less capable of social judgment than humans (Holthöwer & van Doorn, 2022; Kim et al., 2022; Pitardi et al., 2022), interactions with AI provide lower-stakes exposures that allow for easing the user to the idea of outgroup interactions. Provided the interactions are positive, consumers may feel understood, connected, and even empowered (Flavián et al., 2024; Puntoni et al., 2021), reducing the anxiety they anticipate feeling not just when interacting with the AI as a social group proxy member but even with human members of those same groups.

3.3.1 | Example: hate speech moderation

As an example of how these affective processes can be invoked to reduce prejudice, consider the example of hate speech, which is "any kind of communication in speech, writing or behaviour, that attacks or uses pejorative or discriminatory language with reference to a person or a group … based on their religion, ethnicity, nationality, race, colour, descent, gender or other identity factor" (United Nations, 2020). Hate speech is ubiquitous and increasingly happens online and through social media conversations (Parker & Ruths, 2023). Hate speech not only has very adverse emotional, behavioral, and normative consequences for the victim but can also encourage observers of hate speech to respond with contempt instead of empathy to outgroups, further perpetuating hate (Bilewicz & Soral, 2020). Because AI applications like chatbots use machine learning (ML) to learn from human data, they may inadvertently learn hate speech when it is present in the training data, and in turn enact hate speech toward consumers (Caliskan et al., 2017).

Companies are economically incentivized to avoid hate speech on their platforms, yet doing so is challenging given the black-box nature of AI. Because of this, companies have historically taken the "brute force" approach of censoring and avoiding conversations that may involve hate speech, for instance, by employing "blacklists" of forbidden words (Cobbe, 2021) or words often associated with hate speech like "Pakistani" or "black women." Yet, such approaches prioritize caution over precision, since they also sometimes censor messages that have nothing to do with hate speech (Cobbe, 2021) including personal experiences of racism shared by member of marginalized groups (Lee et al., 2024), and they can thus suppress valuable content (Hangartner et al., 2021). Another challenge is that what constitutes hate speech changes over time, meaning that screening mechanisms need to be constantly monitored for accuracy and adapted to account for new expressions of hate speech (Parker & Ruths, 2023).

Companies wishing to meaningfully reduce hate speech in society should not only avoid or suppress it but also provide opportunities for meaningful synthetic contact. For instance, instead of stonewalling users who submit stereotypical inputs, large language models can be finetuned to provide responses that acknowledge nuance and complexity in how to think about group members as well as encourage empathy and calm anxieties around how the user thinks about these groups (Solaiman & Dennison, 2021). By inducing empathy and anxiety-related processes in this targeted way, synthetic contact could lead to prejudice reduction (Bilewicz et al., 2021; Hangartner et al., 2021). Promisingly, recent research finds that, when trying to get Twitter users to reduce or delete racist speech, the most effective interventions are those that underline the negative emotional consequences for people targeted by the hate speech, constituting a clear empathy-based intervention (Hangartner et al., 2021). Similarly, we suggest that inducing empathy in the context of human-AI interactions—as by having the AI exhibit empathy or otherwise prompt the user to engage in empathizing—could be effective at fostering consumer empathy and its social transmission, thereby reducing prejudices and the hate speech it fuels (Chin et al., 2020; Pataranutaporn et al., 2023; Zhou et al., 2024). Furthermore, people have the lay belief that empathy is a limited, zero-sum resource, such that making people believe that empathy is unlimited increases empathy toward outgroups (Hasson et al., 2022). Since AI has an infinite capacity for expressing empathy, empathic AI could naturally foster similar empathy effects. Relatedly, we argue that counter-stereotypical AI that weakens consumers' anxiety about the negative consequences of intergroup contact can have similar effects on hate speech prevention and associated prejudice reduction.

4 | AI-UNIQUE OPPORTUNITIES FOR PREJUDICE REDUCTION

Building on several research literatures—including Computers as Social Actors (Nass et al., 1994; Nass & Moon, 2000), Intergroup Contact Theory (Allport, 1954; Dovidio et al., 2017; Pettigrew & Tropp, 2006), and prejudice reduction (Paluck et al., 2021; Paluck & Green, 2009)—we propose AI-unique opportunities for prejudice reduction through synthetic contact. Relative to contact with human outgroup members, consumers' contact with AI products and services can be more (i) repeated, (ii) direct, (iii) unavoidable, (iv) private, (v) non-judgmental, (vi) collaborative, and (vii) need-satisfying (see Table 1).

4.1 | More repeated

Intergroup Contact Theory research shows that cross-group friendships have special importance in promoting positive contact effects and prejudice reduction, as they typically involve repeated contact

TABLE 1 AI-unique opportunities for prejudice reduction.

AI opportunities		Examples
Interaction type	More Repeated	Contact with AI takes place more repeatedly, frequently, and across more situations.
	More Direct	Contact with AI is direct and personal (e.g., directly asking for advice or receiving recommendations).
	More Collaborative	AI can play an equal agentic role in tasks, cooperatively setting and pursuing goals in tandem with the human user.
Interaction context	More Unavoidable	Contact with AI is more unavoidable since it is often the first and standard option for users to seek advice, complain, interface with firm services, etc.
	More Private	Contact with AI takes place in more private, intimate contexts such as the home, and when users are alone.
Interaction appraisal	More Non-judgmental	AI is perceived as being more neutral and less capable of social judgment.
	More Need-satisfying	AI, being designed for service provision, is more likely to satisfy needs (e.g., provide helpful recommendations, solve customer service problems, provide educational content).

with the outgroup across an extended period and varied settings (Pettigrew, 1998). In fact, a meta-analysis on the effect of cross-group friendships demonstrated that time spent with outgroup friends was the strongest predictor of prejudice reduction (Davies et al., 2011). Yet, intergroup contact with outgroup members is usually limited and often constrained by systems of segregation (Dixon et al., 2005, 2020).

In contrast, consumers' contact with certain AI, such as intelligent personal assistants, typically takes place repeatedly, frequently, and across different contexts, since consumers regularly ask for recommendations, information, and advice (McLean et al., 2021), and AI services are typically easily accessible on demand, 24/7.

4.2 | More direct

Meta-analyses of contact effects demonstrate that direct contact (i.e., face-to-face interactions) with members of outgroups is more effective at reducing prejudices than are forms of indirect contact (Lemmer & Wagner, 2015; Pettigrew & Tropp, 2006). Yet, intergroup contact with outgroup members is typically indirect at best (Dixon et al., 2005)—for instance, people might be exposed to outgroup members through social media, where these outgroups may still be depicted in a stereotyped manner and as segregated from other groups. While indirect contact can sometimes increase intentions to engage in direct contact (e.g., by psychologically preparing humans for direct, face-to-face intergroup contact), these intentions may not necessarily translate into behavior (Ioannou et al., 2018).

Conversely, interactions with AI products and services like intelligent personal assistants and AI companions typically involve direct contact with the technology, and can even be social (Pantano & Scarpi, 2022; Puntoni et al., 2021) and emotional in nature (De Freitas et al., 2024; Yu & Fan, 2024). AI does not just provide opportunities for more intergroup contact of any kind but of direct contact specifically.

4.3 | More unavoidable

Positive contact effects are stronger when individuals cannot avoid contact (aka "no-choice settings;" Pettigrew & Tropp, 2006). However, contact with outgroup members is often avoidable and avoided, because people prefer to associate with others who are similar to themselves, have 'closed' social networks and communities, or simply lack opportunities for contact (Dixon et al., 2005). Avoidance can enable a downward spiral in which prejudices are reinforced—that is, a lack of a positive contact experience creates intergroup anxiety, which in turn reinforces prejudice (Plant & Devine, 2003). Because a lack of contact means that prejudices are not revised, these prejudices continue to prevent further contact, increasing uncertainty and anxiety of intergroup encounters and thereby reinforcing prejudices in turn (Binder et al., 2009).

By comparison, as AI becomes unavoidably integrated into products and services in private, commercial, working, and public contexts, human-AI interactions become increasingly difficult to avoid. Because AI is often a necessary step toward a solution, people may even voluntarily purchase and interact with AI-driven products and services.

4.4 | More private

Contact in the private sphere is likely to be more effective than in the public sphere. In the public sphere, people may be more defensive of their attitudes and ideological positions (Lambert et al., 1996), and more likely to conform to prejudices (Levitan & Verhulst, 2016). Since prejudices are prevalent in the general public and society, people in public settings may reiterate them to socially signal adherence to group norms (Crandall et al., 2002). Although certain types of intergroup contact have become more common, such as cross-group friendships and interracial marriages, their prevalence remains limited due to ongoing social, cultural, and structural barriers (Killen et al., 2022; Pew Research Center, 2017). As a result, people rarely interact with outgroup members (as compared to ingroup members) in private spheres like the home.

In contrast, consumers usually interact with certain AI applications like intelligent personal assistants in the private sphere (McLean & Osei-Frimpong, 2019), including when they are alone at home or more intimate social get-togethers, thereby circumventing the social influences that typically impede public contact from effectively reducing prejudice.

4.5 | More non-judgmental

The concern or anticipation of being negatively evaluated (i.e., intergroup anxiety) is one main barriers to both intergroup contact and prejudice reduction (Plant & Devine, 2003; Stephan, 2014). Firstly, ingroup members may be concerned about being perceived as prejudiced by outgroup members, and about making mistakes or offending them during a first encounter. Second, contact with outgroup members may threaten ingroup members' social identities and distinctiveness and increase their expectation that they will be disapproved by other ingroup members due to social norms against contact (Stephan, 2014).

AI is generally viewed as being less capable of social judgment than humans (Holthöwer & van Doorn, 2022; Kim et al., 2022; Pitardi et al., 2021). Thus, interactions with AI provide less socially judgmental-seeming exposures that can ease the user into the idea of outgroup interactions. Consumers' interactions with counter-stereotypical AI can mitigate people's intergroup anxiety, as users become increasingly familiar and comfortable with the outgroup in low-judgment settings. Synthetic contact with AI is also not associated with the same risks of social identity threat and social norm violations, as compared to human contact.

4.6 | More collaborative

Stronger prejudice reduction occurs when the following conditions are met: (1) equal status between the groups, (2) common goals, (3) intergroup cooperation, and (4) institutional support, i.e., when positive intergroup interactions are endorsed by institutions or social norms (Allport, 1954). Yet there is often a substantial gap between the idealized forms of contact studied by social psychologists and the everyday interactions that characterize most normal encounters between groups (Dixon et al., 2005). Equal status is undermined by socio-economic inequalities and discrimination. The pursuit of common goals is hindered by divergent interests and brief, superficial interactions. And cooperation between groups is weakened by competitive environments in education and workplaces. These forces conspire to impede the ideal conditions for prejudice reduction in real-world settings.

In contrast, AI systems have a better chance of meeting all four conditions. Since providers that offer counter-stereotypical AI thereby endorse their products and services, the fourth condition is fulfilled through the very act of offering counter-stereotypical AI products or services to users. The other three conditions can be met for several types of AI that play a more collaborative role. Consider, for example, the use of AI in education. Some of the latest educational bots, like Kahn Academy's Khanmigo, do not simply answer user queries but socratically guide them toward learning insights. The AI thereby plays an equal agentic role in the learning process, setting intermediate and long-term goals with the learner, and collaborating with them toward achieving those goals (conditions 1–3). Similarly, there are various commercial software applications labeled 'copilots'—such as those offered by Microsoft, Semrush, Salesforce, and Ford—that perform similar collaborative roles, making these interactions conducive to prejudice reduction.

4.7 | More need-satisfying

The final opportunity relates to consumers' need satisfaction. Need satisfaction has been shown to shape human perceptions of social interactions, and to contribute toward the success of close relationships and conflict reduction between groups (Kreienkamp et al., 2023; Paolini et al., 2024). Thus, positive contact effects and prejudice reduction are more likely to arise and last when needs are satisfied (Hässler et al., 2022; Kreienkamp et al., 2023). However, intergroup interactions often do not fulfill the personal and social needs of the people involved (e.g., belongingness, affirmation, respect, empowerment), undermining positive contact effects (Dovidio et al., 2017). This failure to meet needs can stem from several countervailing forces, including power and socio-economic imbalances, preexisting biases, lack of shared experiences, competitive environments, and short-term interactions. In short, several barriers prevent positive and lasting intergroup interactions.

In contrast, AI systems can fulfill a variety of cognitive and affective consumer needs (Hermann, 2022; Huang & Rust, 2024; Puntoni et al., 2021). AI is typically designed and deployed in a human-centric, need-aware manner, given AI providers' and companies' natural motivation to satisfy customers and thereby attract and retain them (Human & Watkins, 2023). Consumer need satisfaction by AI leads to user satisfaction (Xie et al., 2024), which may positively shape the experience and appraisals of AI. Likewise, greater need satisfaction by AI could translate into more effective prejudice reduction.

5 | IMPLEMENTATION CONSIDERATIONS

For counter-stereotypical AI to have its intended societal effects and not inadvertently reproduce and perpetuate prejudices and intergroup biases, it is important that it be implemented in a manner that is mindful of subtyping and tokenism, timing, heterogeneous populations, and preservation of free speech.

5.1 | Subtyping and tokenism

AI providers aiming to counter prejudice through AI need to establish counter-stereotypicality by design across all of their AI products and services, instead of equipping just single products and services with counter-stereotypical attributes. Thereby, AI providers avoid 'subtyping', wherein people perceive single counter-stereotypical AI products or services to be exceptions to the rule, and dissociate these instances from the broader category of products and services (Hutter & Crisp, 2005; Kunda & Oleson, 1995). For instance, a warm and friendly male customer greeting robot will stand out and could be considered an exception if all other friendly customer service robots stereotypically convey the female gender.

A related undesirable outcome is perceived 'tokenism', in which certain stereotyped groups like Black and Brown characters are seen as existing only to support the role of White protagonists (Podoshen et al., 2021). By contrast, implementing consistent counter-stereotypical design across all offerings provides a meaningful representation of social groups and enhances 'spill-over effects', wherein prejudice reduction for one instance generalizes readily to other offerings within the portfolio instead of being compartmentalized (Tausch et al., 2010).

5.2 | Timing

Implementation timing can affect the AI's impact as well. First, an early implementation of counter-stereotypical AI allows for the longer, repeated contact that is needed for prejudices to be revised. However, early implementation of counter-stereotypical AI could exacerbate prejudices if the technology is insufficiently ripe or tested to satisfy needs and create positive contact. AI providers should be particularly careful in ensuring that counter-stereotypical AI is fail-proof (as via thorough prototyping) before implementing it at scale, given that such AI is more likely to be noticed, scrutinized, and moralized by consumers.

Second, first movers on a societal issue like prejudice can appear genuinely committed to the issue, while followers can be perceived as less committed (Nam et al., 2023; Silver et al., 2021). Even so, AI providers may have good reasons for delaying the implementation of counter-stereotypical AI, as when a given social category is highly divisive in culture. In these situations, they can provide a response now in which they promise a well-measured action plan later, thereby showing that they are responsive enough to care while increasing perceptions of effort (Jago & Laurin, 2019) and ethicality (De Kerviler et al., 2022).

5.3 | Heterogenous consumer groups

Counter-stereotypical examples work by both revising stereotypes (the cognitive processes) and emotional reactions (the affective processes). Yet, consumers may have different predispositions toward cognitive and affective processing styless (Cacioppo & Petty, 1982; Maio & Esses, 2001). For instance, some consumers may hold stereotypes even while evincing positive emotions and favorability toward outgroups, whereas others may not hold stereotypes but show negative emotions and unfavorability.

Given this potential heterogeneity and the challenge of foreseeing which types of prejudice are most likely in any given encounter, the best 'blanket approach' is for counter-stereotypical AI to engage both types of processes. For example, a female customer service bot should answer technical questions in both a competent and empathetic way, whereas a black male AI tutor should be both highly knowledgeable and approachable. In these ways, AI not only leads to cognitive recategorization of the underlying stereotypes but does so while dampening intergroup anxiety and elevating empathy.

5.4 | Preservation of free speech

A final significant ethical concern with the use of AI to counteract stereotypes is the potential of efforts to end in a 'slippery slope' toward imposition on freedom of speech. Counter-stereotypical AI should not go so far as to suppress dissenting or traditional opinions or coerce uniformity of thought, especially since (i) many social issues may have nuances that AI systems may not anticipate, (ii) minority voices may be easily marginalized (Mohamed et al., 2020), and (iii) different cultures may have different values that may conflict with the intended values of those who design counter-stereotypical AI. Maintaining the balance between counter-stereotypicality and openness is crucial for protecting freedom of speech, which includes the right to voice perspectives (even those that are stereotypical or unpopular), as long as they do not incite harm or discrimination. Likewise, providers and platforms should strive to preserve the rights of countries to choose the values that they uphold, provided these do not violate internationally agreed-upon human rights.

6 | CONCLUSION

Since "the world has no shortage of challenges to address" (Chandy et al., 2021, p. 7) – prejudices including gender stereotypes and hate speech among them – consumer psychologists and practitioners should make their contributions toward addressing these challenges.

In this context, psychologically informed AI design offers unprecedented opportunities to attain socially good outcomes and enhance social welfare (Morewedge et al., 2023; Valenzuela et al., 2024). Here, we illustrate the opportunity for reducing the cognitive and affective drivers of prejudice by using counter-stereotypical AI, explain why this technology is uniquely suited for this purpose, and provide implementation considerations for achieving this goal effectively.

Zooming out, more than 50 years ago, Kotler and Zaltman (1971) raised the question whether marketing concepts and techniques could be effectively applied to promote social objectives. Since then, social marketing has achieved many successes (Kotler, 2011). The current work takes inspiration from the view that new technology, in this case, AI, can also turbocharge tackling social objectives—especially those that are stubbornly resistant to human interventions alone, like societal prejudice. As with all grand challenges, there are no free or effortless solutions—counter-stereotypical AI will sometimes imply difficult short-term trade-offs between "doing well" and "doing good." Yet, we are convinced that mindful and careful implementation can soften these tradeoffs.

ORCID

Stefano Puntoni https://orcid.org/0000-0002-3259-2325

REFERENCES

Agarwal, S., De Freitas, J., Ragnhildstveit, A., & Morewedge, C. K. (2024). Acceptance of automated vehicles is lower for self than others. *Journal of the Association for Consumer Research*, *9*(3), 269–281. https://doi.org/10.1086/729900

Ahn, J., Kim, J., & Sung, Y. (2022). The effect of gender stereotypes on artificial intelligence recommendations. *Journal of Business Research*, *141*, 50–59. https://doi.org/10.1016/j.jbusres.2021.12.007

Allport, G. W. (1954). *The nature of prejudice*. Addison-Wesley.

Batson, C. D., Polycarpou, M. R., Harmon-Jones, E., Imhoff, H. J., Mitchener, E. C., Bednar, L. L., Klein, T. R., & Highberger, L. (1997). Empathy and attitudes: Can feeling for a member of a stigmatized group improve feelings toward the group? *Journal of Personality and Social Psychology*, *72*(1), 105–118. https://doi.org/10.1037/0022-3514.72.1.105

Berger, B., Adam, M., Rühr, A., & Benlian, A. (2021). Watch me improve: Algorithm aversion and demonstrating the ability to learn. *Business & Information Systems Engineering*, *63*(1), 55–68. https://doi.org/10.1007/s12599-020-00678-5

Bergner, A. S., Hildebrand, C., & Häubl, G. (2023). Machine talk: How verbal embodiment in conversational AI shapes consumer–brand relationships. *Journal of Consumer Research*, *50*(4), 742–764. https://doi.org/10.1093/jcr/ucad014

Bernstein, R. S., Bulger, M., Salipante, P., & Weisinger, J. Y. (2020). From diversity to inclusion to equity: A theory of generative interactions. *Journal of Business Ethics*, *167*(3), 395–410. https://doi.org/10.1007/s10551-019-04180-1

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

Bilewicz, M., & Soral, W. (2020). Hate speech epidemic: The dynamic effects of derogatory language on intergroup relations and political radicalization. *Political Psychology*, *41*(Suppl 1), 3–33. https://doi.org/10.1111/pops.12670

Bilewicz, M., Tempska, P., Leliwa, G., Dowgiałło, M., Tańska, M., Urbaniak, R., & Wroczyński, M. (2021). Artificial intelligence against hate: Intervention reducing verbal aggression in the social network environment. *Aggressive Behavior*, *47*(3), 260–266. https://doi.org/10.1002/ab.21948

Binder, J., Zagefka, H., Brown, R., Funke, F., Kessler, T., Mummendey, A., & Leyens, J.-P. (2009). Does contact reduce prejudice or does prejudice reduce contact? A longitudinal test of the contact hypothesis among majority and minority groups in three European countries. *Journal of Personality and Social Psychology*, *96*(4), 843–856. https://doi.org/10.1037/a0013470

Blut, M., Wang, C., Wünderlich, N. V., & Brock, C. (2021). Understanding anthropomorphism in service provision: A meta-analysis of physical robots, chatbots, and other AI. *Journal of the Academy of Marketing Science*, *49*(4), 632–658. https://doi.org/10.1007/s11747-020-00762-y

Borau, S. (2024). Deception, discrimination, and objectification: Ethical issues of female AI agents. *Journal of Business Ethics*. Advance online publication. https://doi.org/10.1007/s10551-024-05754-4

Borau, S., Otterbring, T., Laporte, S., & Wamba, S. F. (2021). The most human bot: Female gendering increases humanness perceptions of bots and acceptance of AI. *Psychology & Marketing*, *38*(7), 1052–1068. https://doi.org/10.1002/mar.21480

Brough, J., Harris, L. T., Wu, S. H., Branigan, H. P., & Rabagliati, H. (2024). Cognitive causes of “like me” race and gender biases in human language production. *Nature Human Behaviour*. Advance online publication, *8*, 1706–1715. https://doi.org/10.1038/s41562-024-01943-3

Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology*, *42*(1), 116–131. https://doi.org/10.1037/0022-3514.42.1.116

Cadario, R., Longoni, C., & Morewedge, C. K. (2021). Understanding, explaining, and utilizing medical artificial intelligence. *Nature Human Behavior*, *5*(12), 1636–1642. https://doi.org/10.1038/s41562-021-01146-0

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186. https://doi.org/10.1126/science.aal4230

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

Chandy, R. K., Johar, G. V., Moorman, C., & Roberts, J. H. (2021). Better marketing for a better world. *Journal of Marketing*, *85*(3), 1–9. https://doi.org/10.1177/00222429211003690

Chang, E. H., Milkman, K. L., Gromet, D. M., Rebele, R. W., Massey, C., Duckworth, A. L., & Grant, A. M. (2019). The mixed effects of online diversity training. *Proceedings of the National Academy of Sciences*, *116*(16), 7778–7783. https://doi.org/10.1073/pnas.1816076116

Chen, C., Tian, A. D., & Jiang, R. (2023). When post hoc explanation knocks: Consumer responses to explainable AI recommendations. *Journal of Interactive Marketing*, *59*(3), 234–250. https://doi.org/10.1177/10949968231200221

Chin, H., Molefi, L. W., & Yi, M. Y. (2020). Empathy is all you need: How a conversational agent should respond to verbal abuse. In *Proceedings of the 2020 CHI conference on Human factors in computing systems* (pp. 1–13). ACM.

Clegg, M., Hofstetter, R., de Bellis, E., & Schmitt, B. H. (2023). Unveiling the mind of the machine. *Journal of Consumer Research*, *51*(2), 342–361. https://doi.org/10.1093/jcr/ucad075

Cobbe, J. (2021). Algorithmic censorship by social platforms: Power and resistance. *Philosophy & Technology*, *34*(4), 739–766. https://doi.org/10.1007/s13347-020-00429-0

Crandall, C. S., Eshleman, A., & O'Brien, L. (2002). Social norms and the expression and suppression of prejudice: The struggle for internalization. *Journal of Personality and Social Psychology*, *82*(3), 359–378. https://doi.org/10.1037/0022-3514.82.3.359

Crisp, R. J., & Turner, R. N. (2011). Cognitive adaptation to the experience of social and cultural diversity. *Psychological Bulletin*, *137*(2), 242–266. https://doi.org/10.1037/a0021840

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

Davies, K., Tropp, L. R., Aron, A., Pettigrew, T. F., & Wright, S. C. (2011). Cross-group friendships and intergroup attitudes: A meta-analytic review. *Personality and Social Psychology Review*, *15*(4), 332–351. https://doi.org/10.1177/1088868311411103

Davis, M. H. (1994). *Empathy: A social psychological approach*. Brown and Benchmark.

Davis, N., Olsen, N., Perry, V. G., Stewart, M. M., & White, T. B. (2023). I'm only human? The role of racial stereotypes, humanness, and satisfaction in transactions with anthropomorphic sales bots. *Journal of the Association for Consumer Research*, *8*(1), 47–58. https://doi.org/10.1086/722703

De Freitas, J., Agarwal, S., Schmitt, B., & Haslam, N. (2023). Psychological factors underlying attitudes toward AI tools. *Nature Human Behaviour*, *7*(11), 1845–1854. https://doi.org/10.1038/s41562-023-01734-2

De Freitas, J. A. K. U., Uğuralp, Z., & Puntoni, S. (2024). Chatbots and mental health: Insights into the safety of generative AI. *Journal of Consumer Psychology*, *34*(3), 481–491. https://doi.org/10.1002/jcpy.1393

De Kerviler, G., Heuvinck, N., & Gentina, E. (2022). Make an effort and show me the love! Effects of indexical and iconic authenticity on perceived brand ethicality. *Journal of Business Ethics*, *179*(1), 89–110. https://doi.org/10.1007/s10551-021-04779-3

Dietvorst, B. J., & Bartels, D. M. (2022). Consumers object to algorithms making morally relevant tradeoffs because of algorithms' consequentialist decision strategies. *Journal of Consumer Psychology*, *32*(3), 406–424. https://doi.org/10.1002/jcpy.1266

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, *144*(1), 114–126. https://doi.org/10.1037/xge0000033

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2018). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science*, *64*(3), 1155–1170. https://doi.org/10.1287/mnsc.2016.2643

Dixon, J., Durrheim, K., & Tredoux, C. (2005). Beyond the optimal contact strategy: A reality check for the contact hypothesis. *American Psychologist*, *60*(7), 697–711. https://doi.org/10.1037/0003-066X.60.7.697

Dixon, J., Tredoux, C., Davies, G., Huck, J., Hocking, B., Sturgeon, B., Whyatt, D., Jarman, N., & Bryan, D. (2020). Parallel lives: Intergroup contact, threat, and the segregation of everyday activity spaces. *Journal of Personality and Social Psychology*, *118*(3), 457–480. https://doi.org/10.1037/pspi0000191

Dobbin, F., & Kalev, A. (2018). Why doesn't diversity training work? The challenge for industry and academia. *Anthropology Now*, *10*(2), 48–55. https://doi.org/10.1080/19428200.2018.1493182

Dovidio, J. F., Love, A., Schellhaas, F. M. H., & Hewstone, M. (2017). Reducing intergroup bias through intergroup contact: Twenty years of progress and future directions. *Group Processes & Intergroup Relations*, *20*(5), 606–620. https://doi.org/10.1177/1368430217712052

Duani, N., Barasch, A., & Morwitz, V. (2024). Demographic pricing in the digital age: Assessing fairness perceptions in algorithmic versus human-based price discrimination. *Journal of the Association for Consumer Research*, *9*(3), 257–268. https://doi.org/10.1086/729440

Esmaeilzadeh, H., & Vaezi, R. (2022). Conscious empathic AI in service. *Journal of Service Research*, *25*(4), 549–564. https://doi.org/10.1177/10946705221103531

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

Flavián, C., Belk, R. B., Belanche, D., & Casaló, L. V. (2024). Automated social presence in AI: Avoiding consumer psychological tensions to improve service value. *Journal of Business Research*, *175*, 114545. https://doi.org/10.1016/j.jbusres.2024.114545

Fortunati, L., Edwards, A., Edwards, C., Manganelli, A. M., & de Luca, F. (2022). Is Alexa female, male, or neutral? A cross-national and cross-gender comparison of perceptions of Alexa's gender and status as a communicator. *Computers in Human Behavior*, *137*, 107426. https://doi.org/10.1016/j.chb.2022.107426

Garvey, A. M., Kim, T., & Duhachek, A. (2023). Bad news? Send an AI. Good news? Send a human. *Journal of Marketing*, *87*(1), 10–25. https://doi.org/10.1177/00222429211066972

Giroux, M., Kim, J., Lee, J. C., & Park, J. (2022). Artificial intelligence and declined guilt: Retailing morality comparison between human and AI. *Journal of Business Ethics*, *178*(4), 1027–1041. https://doi.org/10.1007/s10551-022-05056-7

Hangartner, D., Gennaro, G., Alasiri, S., Bahrich, N., Bornhoft, A., Boucher, J., & Donnay, K. (2021). Empathy-based counterspeech can reduce racist hate speech in a social media field experiment. *Proceedings of the National Academy of Sciences*, *118*(50), e2116310118. https://doi.org/10.1073/pnas.2116310118

Hässler, T., Ullrich, J., Sebben, S., Shnabel, N., Bernardino, M., Valdenegro, D., & Pistella, J. (2022). Need satisfaction in intergroup contact: A multinational study of pathways toward social change. *Journal of Personality and Social Psychology*, *122*(4), 634–658. https://doi.org/10.1037/pspi0000365

Hasson, Y., Amir, E., Sobol-Sarag, D., Tamir, M., & Halperin, E. (2022). Using performance art to promote intergroup prosociality by cultivating the belief that empathy is unlimited. *Nature Communications*, *13*, 7786. https://doi.org/10.1038/s41467-022-35235-z

Hermann, E. (2022). Anthropomorphized artificial intelligence, attachment, and consumer behavior. *Marketing Letters*, *33*(1), 157–162. https://doi.org/10.1007/s11002-021-09587-3

Hildebrand, C., & Bergner, A. S. (2021). Conversational robo advisors as surrogates of trust: Onboarding experience, firm perception, and consumer financial decision making. *Journal of the Academy of Marketing Science*, *49*(4), 659–676. https://doi.org/10.1007/s11747-020-00753-z

Hilton, J. L., & von Hippel, W. (1996). Stereotypes. *Annual Review of Psychology*, *47*, 237–271. https://doi.org/10.1146/annurev.psych.47.1.237

Hodson, G., Crisp, R. J., Meleady, R., & Earle, M. (2018). Intergroup contact as an agent of cognitive liberalization. *Perspectives on Psychological Science*, *13*(5), 523–548. https://doi.org/10.1177/1745691617752324

Holthöwer, J., & van Doorn, J. (2022). Robots do not judge: Service robots can alleviate embarrassment in service encounters. *Journal of the Academy of Marketing Science*, *51*(4), 767–784. https://doi.org/10.1007/s11747-022-00862-x

Howe, L. C., & Krosnik, J. A. (2017). Attitude strength. *Annual Review of Psychology*, *68*, 327–351. https://doi.org/10.1146/annurev-psych-122414-033600

Huang, M.-H., & Rust, R. T. (2024). The caring machine: Feeling AI for customer care. *Journal of Marketing*, *88*(5), 1–23. https://doi.org/10.1177/00222429231224748

Human, S., & Watkins, R. (2023). Needs and artificial intelligence. *AI and Ethics*, *3*(3), 811–826. https://doi.org/10.1007/s43681-022-00206-z

Hutter, R. R. C., & Crisp, R. J. (2005). The composition of category conjunctions. *Personality and Social Psychology Bulletin*, *31*(5), 647–657. https://doi.org/10.1177/0146167204271575

Hutter, R. R. C., Crisp, R. J., Humphreys, G. W., Waters, G. M., & Moffitt, G. (2009). The dynamics of category conjunctions. *Group Processes & Intergroup Relations*, *12*(5), 673–686. https://doi.org/10.1177/1368430209337471

Ioannou, M., Al Ramiah, A., & Hewstone, M. (2018). An experimental comparison of direct and indirect intergroup contact. *Journal of Experimental Social Psychology*, *76*, 393–403. https://doi.org/10.1016/j.jesp.2017.11.010

Jackson, J. C., Castelo, N., & Gray, K. (2020). Could a rising robot workforce make humans less prejudiced? *American Psychologist*, *75*(7), 969–982. https://doi.org/10.1037/amp0000582

Jago, A. S. (2019). Algorithms and authenticity. *Academy of Management Discoveries*, *5*(1), 38–56. https://doi.org/10.5465/amd.2017.0002

Jago, A. S., Carroll, G. R., & Lin, M. (2022). Generating authenticity in automated work. *Journal of Experimental Psychology: Applied*, *28*(1), 52–70. https://doi.org/10.1037/xap0000365

Jago, A. S., & Laurin, K. (2019). Inferring commitment from rates of organizational transition. *Management Science*, *65*(6), 2842–2857. https://doi.org/10.1287/mnsc.2017.2980

Killen, M., Raz, K. L., & Graham, S. (2022). Reducing prejudice through promoting cross-group friendships. *Review of General Psychology*, *26*(3), 361–376. https://doi.org/10.1177/10892680211061262

Kim, T. W., Jiang, L., Duhachek, A., Lee, H., & Garvey, A. (2022). Do you mind if I ask you a personal question? How AI service agents alter consumer self-disclosure. *Journal of Service Research*, *25*(4), 649–666. https://doi.org/10.1177/10946705221120232

Kipnis, E., Demangeot, C., Pullig, C., Cross, S. N. N., Cui, C. C., Galalae, C., & Williams, J. D. (2021). Institutionalizing diversity-and-inclusion-engaged marketing for multicultural marketplace well-being. *Journal of Public Policy & Marketing*, *40*(2), 143–164. https://doi.org/10.1177/0743915620975415

Kotler, P. (2011). Reinventing marketing to manage the environmental imperative. *Journal of Marketing*, *75*(4), 132–135. https://doi.org/10.1509/jmkg.75.4.132

Kotler, P., & Zaltman, G. (1971). Social marketing: An approach to planned social change. *Journal of Marketing*, *35*(3), 3–12. https://doi.org/10.1177/002224297103500302

Kraus, M. W., Torrez, B., & Hollie, L. (2022). How narratives of racial progress create barriers to diversity, equity, and inclusion in organizations. *Current Opinion in Psychology*, *43*, 108–113. https://doi.org/10.1016/j.copsyc.2021.06.022

Kreienkamp, J., Agostini, M., Bringmann, L. F., de Jonge, P., & Epstude, K. (2023). Need fulfillment during intergroup contact: Three experience sampling studies. *Personality and Social Psychology Bulletin*. Advance online publication. https://doi.org/10.1177/01461672231204063

Kunda, Z., Miller, D. T., & Claire, T. (1990). Combining social concepts: The role of causal reasoning. *Cognitive Science*, *14*(4), 551–577. https://doi.org/10.1207/s15516709cog1404_3

Kunda, Z., & Oleson, K. C. (1995). Maintaining stereotypes in the face of disconfirmation: Constructing grounds for subtyping deviants. *Journal of Personality and Social Psychology*, *68*(4), 565–579. https://doi.org/10.1037/0022-3514.68.4.565

Lambert, A. J., Cronen, S., Chasteen, A. L., & Lickel, B. (1996). Private vs public expressions of racial prejudice. *Journal of Experimental Social Psychology*, *32*(5), 437–459. https://doi.org/10.1006/jesp.1996.0020

Lambrecht, A., & Tucker, C. (2019). Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of STEM career ads. *Management Science*, *65*(7), 2966–2981. https://doi.org/10.1287/mnsc.2018.3093

Lawson, M. A., Martin, A. E., Huda, I., & Matz, S. C. (2022). Hiring women into senior leadership positions is associated with a reduction in gender stereotypes in organizational language. *Proceedings of the National Academy of Sciences*, *119*(9), e2026443119. https://doi.org/10.1073/pnas.2026443119

Lee, C., Gligorić, K., Kalluri, P. R., Harrington, M., Durmus, E., Sanchez, K. L., … Eberhardt, J. (2024). People who share encounters with racism are silenced online by humans and machines, but a guideline-reframing intervention holds promise. *Proceedings of the National Academy of Sciences*, *121*(38), e2322764121. https://doi.org/10.1073/pnas.2322764121

Lemmer, G., & Wagner, U. (2015). Can we really reduce ethnic prejudice outside the lab? A meta-analysis of direct and indirect contact interventions. *European Journal of Social Psychology*, *45*(2), 152–168.

Levitan, L. C., & Verhulst, B. (2016). Conformity in groups: The effects of others' views on expressed attitudes and attitude change. *Political Behavior*, *38*, 277–315. https://doi.org/10.1007/s11109-015-9312-x

Liu-Thompkins, Y., Okazaki, S., & Li, H. (2022). Artificial empathy in marketing interactions: Bridging the human-AI gap in affective and social customer experience. *Journal of the Academy of Marketing Science*, *50*(6), 1198–1218. https://doi.org/10.1007/s11747-022-00892-5

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

Longoni, C., & Cian, L. (2022). Artificial intelligence in utilitarian vs. hedonic contexts: The "word-of-machine" effect. *Journal of Marketing*, *86*(1), 91–108. https://doi.org/10.1177/0022242920957347

Maio, G. R., & Esses, V. M. (2001). The need for affect: Individual differences in the motivation to approach or avoid emotions. *Journal of Personality*, *69*(4), 583–615. https://doi.org/10.1111/1467-6494.694156

Martin, A. E., & Mason, M. F. (2023). Hey Siri, I love you: People feel more attached to gendered technology. *Journal of Experimental Social Psychology*, *104*, 104402. https://doi.org/10.1016/j.jesp.2022.104402

Martin, A. E., & Slepian, M. L. (2021). 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

McLean, G., & Osei-Frimpong, K. (2019). Hey Alexa … examine the variables influencing the use of artificial intelligent in-home voice assistants. *Computers in Human Behavior*, *99*, 28–37. https://doi.org/10.1016/j.chb.2019.05.009

McLean, G., Osei-Frimpong, K., & Barhorst, J. (2021). Alexa, do voice assistants influence consumer brand engagement? – Examining the role of AI-powered voice assistants in influencing consumer brand engagement. *Journal of Business Research*, *124*, 312–328. https://doi.org/10.1016/j.jbusres.2020.11.045

Mohamed, S., Png, M.-T., & Isaac, W. (2020). Decolonial AI: Decolonial theory as sociotechnical foresight in artificial intelligence. *Philosophy & Technology*, *33*(4), 659–684. https://doi.org/10.1007/s13347-020-00405-8

Morewedge, C. K., Mullainathan, S., Naushan, H. F., Sunstein, C. R., Kleinberg, J., Raghavan, M., & Ludwig, J. O. (2023). Human bias in algorithm design. *Nature Human Behaviour*, *7*(11), 1822–1824. https://doi.org/10.1038/s41562-023-01724-4

Nam, J., Balakrishnan, M., De Freitas, J., & Brooks, A. W. (2023). Speedy activists: How firm response time to sociopolitical events influences consumer behavior. *Journal of Consumer Psychology*, *33*(4), 632–644. https://doi.org/10.1002/jcpy.1380

Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, *56*(1), 81–103. https://doi.org/10.1111/0022-4537.00153

Nass, C., Steuer, J., Henriksen, L., & Dryer, D. C. (1994). Machines, social attributions, and ethopoeia: Performance assessments of computers subsequent to "self-" or "other-" evaluations. *International Journal of Human-Computer Studies*, *40*(3), 543–559. https://doi.org/10.1006/ijhc.1994.1025

Nielsen, Y. A., Pfattheicher, S., & Keijsers, M. (2022). Prosocial behavior toward machines. *Current Opinion in Psychology*, *43*, 260–265. https://doi.org/10.1016/j.copsyc.2021.08.004

Paluck, E. L., & Green, D. P. (2009). Prejudice reduction: What works? A review and assessment of research and practice. *Annual Review of Psychology*, *60*, 339–367. https://doi.org/10.1146/annurev.psych.60.110707.163607

Paluck, E. L., Porat, R., Clark, C. S., & Green, D. P. (2021). Prejudice reduction: Progress and challenges. *Annual Review of Psychology*, *72*, 533–560. https://doi.org/10.1146/annurev-psych-071620-030619

Pantano, E., & Scarpi, D. (2022). I, robot, you, consumer: Measuring artificial intelligence types and their effect on consumers emotions in service. *Journal of Service Research*, *25*(4), 583–600. https://doi.org/10.1177/10946705221103538

Paolini, S., Gibbs, M., Sales, B., Anderson, D., & McIntyre, K. (2024). Negativity bias in intergroup contact: Meta-analytical evidence that bad is stronger than good, especially when people have the opportunity and motivation to opt out of contact. *Psychological Bulletin.*, *150*(8), 921–964. https://doi.org/10.1037/bul0000439

Park, Y. W., Voss, G. B., & Voss, Z. G. (2023). Advancing customer diversity, equity, and inclusion: Measurement, stakeholder influence, and the role of marketing. *Journal of the Academy of Marketing Science*, *51*(1), 174–197. https://doi.org/10.1007/s11747-022-00883-6

Parker, S., & Ruths, D. (2023). Is hate speech detection the solution the world wants? *Proceedings of the National Academy of Sciences*, *120*(10), e2209384120. https://doi.org/10.1073/pnas.2209384120

Pataranutaporn, P., Liu, R., Finn, E., & Maes, P. (2023). Influencing human–AI interaction by priming beliefs about AI can increase perceived trustworthiness, empathy, and effectiveness. *Nature Machine Intelligence*, *5*(10), 1076–1086. https://doi.org/10.1038/s42256-023-00720-7

Pettigrew, T. F. (1998). Intergroup contact theory. *Annual Review of Psychology*, *49*, 65–85. https://doi.org/10.1146/annurev.psych.49.1.65

Pettigrew, T. F. (2009). Secondary transfer effect of contact: Do intergroup contact effects spread to noncontacted outgroups? *Social Psychology*, *40*(2), 55–65. https://doi.org/10.1027/1864-9335.40.2.55

Pettigrew, T. F., & Tropp, L. R. (2006). A meta-analytic test of intergroup contact theory. *Journal of Personality and Social Psychology*, *90*(5), 751–783. https://doi.org/10.1037/0022-3514.90.5.751

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

Pew Research Center. (2017). *Intermarriage in the U.S. 50 years after loving v.* Pew Research Center. https://www.pewresearch.org/social-trends/2017/05/18/intermarriage-in-the-u-s-50-years-after-loving-v-virginia/

Pitardi, V., Wirtz, J., Paluch, S., & Kunz, W. H. (2022). Service robots, agency and embarrassing service encounters. *Journal of Service Management*, *33*(2), 389–414. https://doi.org/10.1108/JOSM-12-2020-0435

Plant, E. A., & Devine, P. G. (2003). The antecedents and implications of intergroup anxiety. *Personality and Social Psychology Bulletin*, *29*(6), 790–801. https://doi.org/10.1177/0146167203029006011

Podoshen, J. S., Ekpo, A. E., & Abiru, O. (2021). Diversity, tokenism, and comic books: Crafting better strategies. *Business Horizons*, *64*(1), 131–140. https://doi.org/10.1016/j.bushor.2020.10.006

Puntoni, S., Reczek, R. W., Giesler, M., & Botti, S. (2021). Consumers and artificial intelligence: An experiential perspective. *Journal of Marketing*, *85*(1), 131–151. https://doi.org/10.1177/0022242920953847

Rathee, S., Banker, S., Mishra, A., & Mishra, H. (2023). Algorithms propagate gender bias in the marketplace—With consumers' cooperation. *Journal of Consumer Psychology*, *33*(4), 621–631. https://doi.org/10.1002/jcpy.1351

Reich, T., Kaju, A., & Maglio, S. J. (2023). How to overcome algorithm aversion: Learning from mistakes. *Journal of Consumer Psychology*, *33*(2), 285–302. https://doi.org/10.1002/jcpy.1313

Schiller, A., & McMahon, J. (2019). Alexa, alert me when the revolution comes: Gender, affect, and labor in the age of home-based artificial intelligence. *New Political Science*, *41*(2), 173–191. https://doi.org/10.1080/07393148.2019.1595288

Silver, I., Kelley, B. A., & Small, D. A. (2021). Selfless first movers and self-interested followers: Order of entry signals purity of motive in pursuit of the greater good. *Journal of Consumer Psychology*, *53*(2), 501–517. https://doi.org/10.1002/jcpy.1228

Solaiman, I., & Dennison, C. (2021). Process for adapting language models to society (palms) with values-targeted datasets. *Advances in Neural Information Processing Systems*, *34*, 5861–5873.

Stephan, W. G. (2014). Intergroup anxiety: Theory, research, and practice. *Personality and Social Psychology Review*, *18*(3), 239–255. https://doi.org/10.1177/1088868314530518

Stephan, W. G., & Finlay, K. (1999). The role of empathy in improving intergroup relations. *Journal of Social Issues*, *55*(4), 729–743. https://doi.org/10.1111/0022-4537.00144

Stephan, W. G., & Stephan, C. W. (1985). Intergroup anxiety. *Journal of Social Issues*, *41*(3), 157–175. https://doi.org/10.1111/j.1540-4560.1985.tb01134.x

Tausch, N., Hewstone, M., Kenworthy, J. B., Psaltis, C., Schmid, K., Popan, J. R., Cairns, E., & Hughes, J. (2010). Secondary transfer effects of intergroup contact: Alternative accounts and underlying processes. *Journal of Personality and Social Psychology*, *99*(2), 282–302. https://doi.org/10.1037/a0018553

Torrez, B., Hollie, L., Richeson, J. A., & Kraus, M. W. (2024). The misperception of organizational racial progress toward diversity, equity, and inclusion. *American Psychologist*, *79*(4), 581–592. https://doi.org/10.1037/amp0001309

Tropp, L. R., & Pettigrew, T. F. (2005). Differential relationships between intergroup contact and affective and cognitive dimensions of prejudice. *Personality and Social Psychology Bulletin*, *31*(8), 1145–1158. https://doi.org/10.1177/0146167205274854

Turner, R. N., Hewstone, M., & Voci, A. (2007). Reducing explicit and implicit outgroup prejudice via direct and extended contact: The mediating role of self-disclosure and intergroup anxiety. *Journal of Personality and Social Psychology*, *93*(3), 369–388. https://doi.org/10.1037/0022-3514.93.3.369

United Nations. (2020). United Nations strategy and plan of action on hate speech. https://www.un.org/en/genocideprevention/hate-speech-strategy.shtml

Uysal, E., Alavi, S., & Bezençon, V. (2022). Trojan horse or useful helper? A relationship perspective on artificial intelligence assistants with humanlike features. *Journal of the Academy of Marketing Science*, *50*(6), 1153–1175. https://doi.org/10.1007/s11747-022-00856-9

Valenzuela, A., Puntoni, S., Hoffman, D., Castelo, N., De Freitas, J., Dietvorst, B., & Wertenbroch, K. (2024). How artificial intelligence constrains the human experience. *Journal of the Association for Consumer Research*, *9*(3), 241–256. https://doi.org/10.1086/730709

Vanman, E. J. (2016). The role of empathy in intergroup relations. *Current Opinion in Psychology*, *11*, 59–63. https://doi.org/10.1016/j.copsyc.2016.06.007

Wien, A. H., & Peluso, A. M. (2021). Influence of human versus AI recommenders: The roles of product type and cognitive processes. *Journal of Business Research*, *137*, 13–27. https://doi.org/10.1016/j.jbusres.2021.08.016

Wölfer, R., Christ, O., Schmid, K., Tausch, N., Buchallik, F. M., Vertovec, S., & Hewstone, M. (2019). Indirect contact predicts direct contact: Longitudinal evidence and the mediating role of intergroup anxiety. *Journal of Personality and Social Psychology*, *116*(2), 277–295. https://doi.org/10.1037/pspi0000146

Woods, H. S. (2018). Asking more of Siri and Alexa: Feminine persona in service of surveillance capitalism. *Critical Studies in Media Communication*, *35*(4), 334–349. https://doi.org/10.1080/15295036.2018.1488082

Xie, C., Wang, Y., & Cheng, Y. (2024). Does artificial intelligence satisfy you? A meta-analysis of user gratification and user satisfaction with AI-powered chatbots. *International Journal of Human–Computer Interaction*, *40*(3), 613–623.

Xie, Z., Yu, Y., Zhang, J., & Chen, M. (2022). The searching artificial intelligence: Consumers show less aversion to algorithm-recommended search product. *Psychology & Marketing*, *39*(10), 1902–1919. https://doi.org/10.1002/mar.21706

Yang, W., & Xie, Y. (2024). Can robots elicit empathy? The effects of social robots' appearance on emotional contagion. *Computers in Human Behavior: Artificial Humans*, *2*(1), 100049.

Yi, A., & Turner, B. (2024). Representations and consequences of race in AI systems. *Current Opinion in Psychology*, *58*, 101831.

Yu, L., & Fan, X. (2024). Lonely human and dominant robot: Similarity versus complementary attraction. *Psychology & Marketing.*, *41*(5), 1133–1151. https://doi.org/10.1002/mar.21975

Zhou, Y., Han, S., Kang, P., Tobler, P. N., & Hein, G. (2024). The social transmission of empathy relies on observational reinforcement learning. *Proceedings of the National Academy of Sciences*, *121*(9), e2313073121. https://doi.org/10.1073/pnas.2313073121

Zhu, Y., Zhang, J., Wu, J., & Liu, Y. (2022). AI is better when I'm sure: The influence of certainty of needs on consumers' acceptance of AI chatbots. *Journal of Business Research*, *150*, 642–652. https://doi.org/10.1016/j.jbusres.2022.06.044

How to cite this article: Hermann, E., De Freitas, J., & Puntoni, S. (2024). Reducing prejudice with counter-stereotypical AI. *Consumer Psychology Review*, 1–12. https://doi.org/10.1002/arcp.1102