Why Some are Better Than Others at Detecting Social Bots: Comparing Baseline Performance to Performance with Aids and Training

Submitted in partial fulfillment of the requirements for the
degree of
Doctor of Philosophy
in
Department of Engineering and Public Policy

LTC Ryan John Kenny

B.A., Psychology, University of Notre Dame
M.A., National Security and Strategic Studies, U.S. Naval War College

Carnegie Mellon University
Pittsburgh, PA

May, 2022

ACKNOWLEDGEMENTS

I want to acknowledge the incredible mentorship, coaching, and support of my advisor Dr. Baruch Fischhoff, and the other members of my committee, Dr. Alex Davis, Dr. Casey Canfield, and Dr. Stephen Broomell. I also appreciate collaborating with Dr. Kathleen Carley and Dr. Dave Beskow in producing this research. I want to thank the Swedish Foundation for Humanities and Social Science for support in funding study 1. Without funding through the Army's Strategic Policy and Planning Program (ASP3) Goodpaster Scholars Program, and the incredible support and professionalism of Dr. Robert Davis and the ASP3 support staff, this intellectual journey would never have occurred. I wish to acknowledge the important roles Dr. Laura Carlson, Dr. Randolph Easton, and Dr. Jeanne Scholl played in my early development as a young researcher and the tough lessons they helped me learn along the way. Without the unconditional love of my dad, Robert Kenny, drive for independence and resourcefulness instilled by my single-mother, Connie Murray, and the abiding kindness and support of Greg and Kathy Murphy, my in-laws, I never would have achieved this accomplishment or likely few others in my life. Finally, and most importantly, without the love, dedication, and patience of my wife, Megan Kenny, and the constant encouragement of our children, John and Mary Kenny, none of this would have happened, let alone mattered. They are the best part of my life, and they inspire me every day to better myself.

ABSTRACT

Social bots have infiltrated many social media platforms, sowing misinformation and disinformation. The harm caused by social bots depends on their ability to avoid detection by credibly impersonating human users. These three studies use a signal detection task to compare human detection of Twitter social bot personas with that of machine learning assessments. Across these studies, we find that sensitivity was (1a) minimal without training or aid, (1b) people were hesitant to respond ‘bot,’ and (1c) people were prone to “myside bias,” judging personas less critically when they shared political views. We also observed (1d) sensitivity improved when a bot detection aid was provided and (1e) when users received training focused on the objectives of social bot creators: to amplify narratives to an extensive social network. When participants labeled a persona a social bot, (2) the probability of their willingness to share its content dropped dramatically. We investigated the relationships between users’ attributes and social bot detection performance and found, (3a) social media experience did not improve detection and at times impaired it; (3b) myside bias affected the sensitivity and criterion used by liberals and conservatives differently; and (3c) analytical reasoning did not improve social bot detection, nor did it mitigate observed myside bias effects, but increased them slightly. We found that (4) people were more concerned about social bots influencing others’ online behaviors than being influenced themselves. Additionally, users’ willingness to pay for a social bot detection aid increased (5a) the more they were concerned about social bots, (5b) the greater their social media experience, (5c) the greater their sensitivity, and (5d) the higher their threshold for responding ‘bot.’ These findings demonstrate the threat posed by social bots and two interventions that may reduce them.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1. The Social Bot Problem

Social media users face an increasingly challenging Turing Test when interacting with other users within their social networks. They must decide whether they believe online personas are humans or social bots. Most social bots go unnoticed as they mimic human users (Cresci, 2020; Stieglitz, Brachten, Ross, & Jung, 2017). The majority of online bots execute mundane advertising tasks (Appel, Grewal, Hadi, & Stephen, 2020). However, some engage in harmful activities such as disseminating misinformation and encouraging false beliefs (Cresci, 2020; Shao, Ciampaglia, Varol, Flammini, & Menczer, 2017; Ferrara, Varol, Davis, Menczer & Flammini, 2016; Torres-Lugo, Truong, Wu, Morstatter, Beskow, & Carley, 2020; Carley & Liu, 2019; Huang, & Carley, 2020; Pacheco, Hui, Wang, Angarita, & Renna, 2018).

The expanded use of social bots has motivated the development of detection tools such as Botometer (Davis, Varol, Ferrara, Flammini, & Menczer, 2016; Yang, Ferrara, & Menczer, 2022), Bot Hunter (Beskow, & Carley, 2018a; Beskow, & Carley, 2018b), and Bot Sight (Kats, 2022). In turn, as these tools have demonstrated their value by identifying bots, a cat and mouse game has materialized, with social bot developers who wish to avoid detection. Evidence suggests social bots continue to evolve to better disguise themselves as genuine accounts (Cresci, Petrocchi, Spognardi, & Tognazzi, S., 2021; Cresci, 2020).

As the competition to avoid detection escalates, there is growing interest in protecting online users from social bot deception through improved detection (Cresci, 2020). Some observers have argued bots share valid information as much as they spread false content (Vosoughi, et al., 2018). However, recent evidence suggests that social bots proliferate low credibility information disproportionally in online social networks, potentially influencing socio-political events (Uyheng, Ng, & Carley2021; Hajli, Saeed, Tajvidi, & Shirazi, 2021; Bastos, & Mercea, 2019; Shao, Ciampaglia, Varol, Yang, Flammini & Menczer, 2017; Shorey & Howard, 2016; Bessi, & Ferrara, 2016; Caldarelli, De Nicola, Del Vigna, Petrocchi & Saracco, 2020), spreading hate (Uyheng & Carley, 2020, 2021), and undermining public health (Ferrara, Cresci, & Luceri, 2020; Shahi, Dirkson, & Majchrzak, 2021; Caldarelli, De Nicola, Petrocchi, Pratelli, & Saracco, 2021). Improving social bot detection may lead to less deception and the spread of false information and narratives.

Social Bots on Twitter

Bot activity is notably widespread on Twitter, a social media microblogging and social networking service employing short (280 characters or fewer) messages called tweets (Chu, Gianvecchio, Wang & Jajodia, 2010). In 2018, Twitter disclosed that it had deleted approximately 70 million fake accounts within two months (Timberg & Dwoskin, 2018). Analysts estimate that 9% - 15% of active Twitter accounts are non-human bots (Varol, Ferrara, Davis, Menczer & Flammini, 2017). Social bots targeting socio-political activity on Twitter has been estimated as high as 25% - 30% of all traffic (Huang, 2020; Uyheng & Carley, 2020).

Within Twitter, social bots may be employed to inflate follower counts, generate message “likes,” and induce other users to share, or “retweet,” their content. Some social bots are fully automated, while others have varying human oversight and control (Stieglitz, Brachten, Ross & Jung, 2017). Twitter users seeking large audiences of followers (i.e., want to be *influencers*) can employ social bots to simulate engagement and flood social networks with content while drowning out other voices (Jansen, Zhang, Sobel & Chowdury, 2009; Cook, Waugh, Abdipanah, Hashem & Rahman, 2014; Lee, Kwak, Park & Moon, 2010; Riquelme, & González-Cantergiani, 2016).

Detecting Bot Signals

Evaluating an online persona requires an observer to examine evidence to decide whether to treat a persona as a human or a bot. Signal Detection Theory (SDT) formalizes such tasks in terms of two components: (a) *discriminability* or *sensitivity*, reflecting how well an observer can distinguish the two possible states; and (b) *decision rule* or *criterion*, reflecting an observer’s preferences for making correct judgments and avoiding incorrect ones (Green & Swets, 1966; McNicol, 2005).

In a SDT task (e.g., Diehl, 1981; Liberman, et al., 1957), a person (or instrument) observes a noisy stimulus and decides whether it contains a specific "signal." Early psychological sensory-perceptual tasks involved auditory tones, faint visual lights, and haptic stimuli. Participants' detection ability is tested with variations in these stimuli (Green & Swets, 1966; Macmillan, & Creelman, 2004). More recently, SDT methods have been applied to more complex decision-making tasks involving discrimination between states (Lynn, & Barrett, 2014; Sorkin, & Woods, 1985; Bisseret, 1981; Canfield, Fischhoff, & Davis, 2016).

In studies of deception, SDT methods have been used to assess the factors influencing observers ‘sensitivity and criteria when assessing potentially duplicitous phenomena (Warren-West, & Jackson, 2020; Bond Jr, & DePaulo, 2006; Bond, 2008; Hauch, Sporer, Michael, & Meissner, 2016). Judging whether a persona is a social bot requires the observer to examine indicators in a persona's profile and activity patterns and decide whether the evidence supports treating it as a human or bot, reflecting how the observer weighs the possible outcomes. Because our studies ask participants to look for bots, we treat ‘bot’ characteristics as the ‘signal’ and human characteristics as the noise.

In SDT terms, bot detection can yield four possible outcomes: successfully identifying a social bot, a *hit* (or true positive); classifying a bot as a human, a *miss* (or false negative); identifying a human as a bot, a *false alarm* (or false positive); or placing a human persona as a human, a *correct rejection* (or true negative). SDT uses an observer's hit and miss rate to estimate performance. SDT characterizes performance in terms of two parameters: Sensitivity to differences in stimuli (d') and the decision-making criterion or threshold (c) for acting on beliefs.

In any SDT trial, a stimulus will present either a signal or a signal plus noise. Each type of trial forms the basis of two distinct Gaussian distributions mean-centered for signal and signal plus noise distributions with equal variance amongst these distributions (McNichol, 2005). In SDT tasks, the underlying distributions are not observed; they are inferred from the observed hit, correct rejection, miss, and false alarm rates.

Sensitivity (d') is the standardized difference between an observer's hit and false alarm rates.

$$d' = z(\mathrm{FA}) - z(\mathrm{H})$$

Larger values of d' represent better sensitivity, while values near zero indicate chance performance.

The criterion reflects an observer's willingness to respond "signal" under uncertainty. A criterion where false alarm rates and misses are equal is considered unbiased and resides halfway between the signal and noise distributions. The extent to which one response ("signal" or "noise") is more common than its base rate in the stimulus population is known as *bias*.

A criterion that is more opposed to false alarms (e.g., treating a human as a bot) than to misses (e.g., treating a bot as a human) is typically called "conservative" and the converse "liberal." However, as the studies in this dissertation also consider political conservatism and liberalism, we will use alternative wording. If observers are hesitant to respond 'bot,' we label them as having higher criteria and biased against treating a human as a bot. Conversely, if they readily respond 'bot,' we label them as having a lower criterion and are biased toward responding 'bot.' Other things being equal, the higher the criterion, the more likely personas are to be called humans (including a higher portion of bots). Human and bot stimuli are equally likely in all three studies in the dissertation. Therefore, observers who treat a majority of personas as humans have higher thresholds for responding 'bot.'

Observers vary their criteria based on the perceived consequences of their decision. For example, oncological radiologists examining chest x-rays for the presence of tumors may have a lower criterion for treating an image as a suspicious signal for fear of missing a life-threatening disease (and, conversely, be less stringent about dismissing it as noise). That criterion would increase both their hit and their false alarm rates. Conversely, a mechanic examining a car during a general maintenance inspection may not respond "signal" when she hears an odd sound for owners on a strict budget or has work backlogged. That higher criterion for treating the car needing additional attention could increase correct rejections and misses.

Automated Social Bot Detection Tools

Social bot detection algorithms use features associated with Twitter accounts and their social networks to inform their predictions. They are trained to predict the conditional relations between those features and outcome variables characterized by humans or other algorithms. Once a social bot detection algorithm has been trained and its parameters are known either publicly or can be inferred, an adversary can create new data patterns designed to mask its presence (Kurakin, Goodfellow, & Bengio, 2016; Wiyatno, Xu, Dia, & de Berker, 2019). That dynamic means that even if observers have and can use bot detection algorithms (the topic of Chapter 3), they will still need to rely on judgment (the topic of Chapter 4).

The development of social bot detection algorithms has focused on two main approaches. The first makes use of the predictive features of individual users. Twitter maintains these features as User Objects in Twitter User account metadata. The second approach considers the platform's network structure to identify coordinated bots (Hjouji, Hunter, Mesnards, Zaman, 2018; Ferrara, Varol, Davis, Menczer, Flammini, 2016; Beskow & Carley, 2018c). This information is derived from *Tweet Objects*, their associated metadata, and the metadata of the users that either share or interact with Tweet Objects.

Social Bot Predictive Features of Individual Users

In managing their Twitter accounts, individuals can manipulate some features of their persona. For example, they can craft different personas by varying profile and background images, editing persona descriptions, and deciding what content to retweet and like. These *user-defined* features afford the creators of social bots the best opportunity to disguise their accounts. They can readily devise persona details that match other 'typical' users' profiles and thus increase the noise to signal ratio for their profiles. As a result, social bot detection algorithms increasingly focus on features that are not user-defined but managed by the Twitter platform.

Social Bot Predictive Features Managed by Twitter Platform

Platform-managed information tracks the online behavior of personas. Some of these features are visible to other users, such as a persona's age, their total number of Tweets, and number of personas that it follows and that follow it. Other features are not generally visible, such as when an account tweets and what other accounts interact with it and share its content. Twitter routinely monitors some of these features and removes accounts with patterns that identify them as non-human, such as *amplifier bots* that propagate messages working at great volume and at all times of day and night (Beskow & Carley, 2019).

Human Social Bot Detection

A 2018 Pew Research Center survey of Americans found that most respondents reported being aware of the existence of social bots. However, only half were at least "somewhat confident" that they could identify them, with only 7% being "very confident" (Stocking & Sumida, 2018). If those self-assessments are accurate, many users may follow social bots and unwittingly share their content (Mønsted, Sapieżyński, Ferrara & Lehmann, 2017). However, a gap exists between these self-reports and observed social bot detection performance.

In 2019, when we began this line of research, we could find no empirical investigation of how people deal with social bots. Therefore, we first developed an SDT-based methodology to provide statistical inferences of social media users' sensitivity (d') and decision criteria (c) when responding to Twitter accounts that might represent social bots. Our methodology asks participants to evaluate actual Twitter persona profiles with varying bot signal strengths, as characterized by the bot indicator scores produced by bot detection algorithms. In subsequent chapters, we describe these tasks, detection algorithms, and analytical methods.

Research Agenda

In the following studies, we investigated (1) the baseline ability of individuals to detect social bots, (2) the benefit of real-time bot indicator aids, and (3) the effects of a social bot detection training protocol on improving social bot detection abilities. In answering each of these questions, we consider characteristics of individuals (task engagement, social media experience, analytical reasoning, and political views) and characteristics of stimuli (signal strength of being a social bot, and political differences between stimuli and individual). The final two studies investigate two behavioral responses (4) willingness to share the content of tweets that might represent a social bot and (5) willingness to pay for an automated social bot detection aid.

Chapter 2. Social Bot Detection

There are two mechanisms for detecting social bots: automatic algorithms and human users. Here, we study the latter, using characterizations produced by the former. We evaluate human performance in detecting Twitter social bots in signal detection theory (SDT) terms (Green & Swets, 1966; Macmillan & Creelman, 2004). In SDT terms, bot detection has four possible outcomes: A *hit* (or true positive), successfully identifying a social bot. A *miss* (or false negative), classifying a bot as a human. A *false alarm* (or false positive), identifying a human as a bot. A *correct rejection* (or true negative), identifying a human persona as a human. SDT characterizes a judge's performance in terms of two parameters: *sensitivity* to differences in stimuli (d') and decision-making *criterion* or threshold (c) for acting on beliefs. Here, we estimate those parameters, then examine how they vary with participant and stimulus (persona) characteristics in a simulated social bot detection task.

Our task asks participants to examine Twitter persona profiles and judge whether each is a bot or a human. We asked them to examine persona profiles and not simply tweets because tweets alone do not reveal the persona's characteristics. A tweet contains the persona's name, the body of the message, and popularity details, such as the number of users who 'liked' or 'retweeted' the message. A user who was suspicious about the authenticity of a persona could examine its profile to inform their judgment. We selected Twitter personas that were politically active during the 2018 midterm election. To estimate a persona's probability of being a social bot, we used two machine-learning social bot detection systems for Twitter: Bot-hunter (Beskow, & Carley, 2018) and Botometer (Davis, Varol, Ferrara, Flammini & Menczer, 2016). The two systems were developed independently and trained with different data. Both produce a probability of a persona being a social bot. We combined them to create a *bot-indicator score* for each persona.

Botometer uses six categories of features as evidence for its bot probability score (Davis, Varol, Ferrara, Flammini & Menczer, 2016) (see Figure 1A, Appendix A). *Network features* scores patterns of information diffusion on the Twitter platform. *User features* are account metadata, including language, location, and date created. *Friends features* pertain to an account's social contacts. *Temporal features* catalog when content was created and shared. *Content features* are language cues within tweets. *Sentiment features* capture the emotional tone of an account's tweets using sentiment analysis algorithms. Botometer was initially trained on a data set generated using honeypots to attract social bots (Lee, Eoff, & Caverlee, 2011). It is retrained periodically with new datasets to compensate for drift in bot characteristics over time (Botometer, 2021). The version employed here was accessed in March 2019.

Bot-hunter bases its predictions on tiers of similar features, ranging from account and tweet information (the lowest two tiers) to temporal and network analyses (at the highest tiers) (Beskow & Carley, 2018). The bot indicator scores used here were based on predictions using Bot Hunter's Tier 1 account and tweet information, as most comparable to what average Twitter users encounter. To train its Tier 1 models, Bot-hunter used several legacy datasets (Beskow, 2020), annotated data captured in a bot attack on NATO and the Digital Forensic Labs (Beskow & Carley, 2018), suspended Russian bot data released by Twitter in October 2018 (Twitter, 2019), and suspended accounts gathered for this purpose.

Stimuli were selected so that bot indicator scores were roughly uniformly distributed from very low (1%) to very high (98%), with 2% intervals (no stimuli scored above 98%). We initially selected a set based on Bot-hunter, then eliminated ones where the Botometer score differed by more than 0.1%. There were relatively few stimuli with high Bot-hunter scores (> 85%), so we relaxed the criterion for eight stimuli in that range. Botometer still gave high probabilities for those personas being bots, just lower ones than Bot-hunter. We used the Bot-Hunter probability as the bot indicator score for these stimuli. For all other stimuli, we used the matched probability.

Predicted Relationships

Task Characteristics

Sensitivity. Sensitivity should be greater when bot indicator scores are further from 50%, indicating personas that are more clearly bots or humans.

Criterion. As their performance has no real-world consequences, we expected participants to put equal value on the four possible outcomes (hits, misses, false alarms, correct rejections). If so, and they believe that bot and human stimuli are equally common, they should respond "bot" and "human" equally often (Canfield, Fischhoff, & Davis, 2016). Choosing “bot” more often would suggest an aversion to misidentifying one of them, similarly to responding “human” more often.

Participant Characteristics

Social Media Experience. We expected participants with more social media experience to have greater sensitivity, assuming that experience has provided useful feedback (Langley, 1985). Indeed, Bot-hunter's annotation relies on experts presumed to have such experience (Weiss & Shanteau, 2003; Landy, 2018; Endsley, 2018). We did not expect social media experience to affect participants' decision criteria.

Analytical Reasoning Ability. Individuals' performance depends not just on their knowledge but also on how well they deploy it. Frederick (2005) developed the Cognitive Reflection Test (CRT) to measure the propensity to overcome initial impulses and engage in reflective, analytical reasoning. People with higher CRT scores have been better on discrimination and judgment tasks similar to the present task (Campitelli & Labollita, 2010; Bar-Hillel, Noah & Frederick, 2019; Toplak, West & Stanovich, 2011). Thus, we expected participants with higher CRT scores to have greater sensitivity, reflected in judgments more strongly correlated with the bot indicator score. We had no reason to expect this ability to shift participants' criteria.

Myside Bias. Social media users tend to follow and engage with sources that agree with them (Flaxman, Goel, & Rao, 2016; Stroud, 2008; Bakshy, Messing & Adamic, 2015). However, even within media bubbles, not all messages confirm existing beliefs. Myside bias is the tendency to examine messages less critically if they support one's political views (Stanovich, West & Toplak, 2013; Stanovich, & West, 2008; Toplak, & Stanovich, 2003; Drummond & Fischhoff, 2019). It has been found to affect how people evaluate evidence, generate arguments, and test hypotheses, in varied settings, including how people evaluate online information sources and share messages (Westerwick, Johnson & Knobloch-Westerwick, 2017; Barberá, Jost, Nagler, Tucker, & Bonneau, 2015).

We examined myside bias in terms of how participants responded to messages varying in the correspondence between their self-reported political alignment and that of the Twitter personas. For the latter, two judges (RK and his wife MK) independently scored each persona as “conservative,” “moderate,” or “liberal.” The two judges discussed and reconciled any differences (occurring with 5 of the 50 profiles). We expected myside bias to reduce performance, reflected in lower sensitivity when messages shared participants’ political orientation. We also expected a criterion shift, with participants more willing to believe that "myside" personas were humans.

Controls. We included two control variables in our prediction models: (a) *Stimulus presentation order* to see if fatigue reduced sensitivity later in the study (Parasuraman & Davies, 1977) and (b) *Task engagement* to see if more engaged participants performed differently. We assessed engagement with one attention check following the instructions and two randomly embedded in the experimental tasks. We expected participants who answered more attention checks correctly to demonstrate greater sensitivity (Matthews, Warm, Reinerman, Langheim & Saxby, 2010; Downs, Holbrook, Sheng, & Cranor, 2010; Dewitt, Fischhoff, Davis & Broomell, 2015). We expected no correlation between either control variable and participants' decision criteria.

Methods

Sample

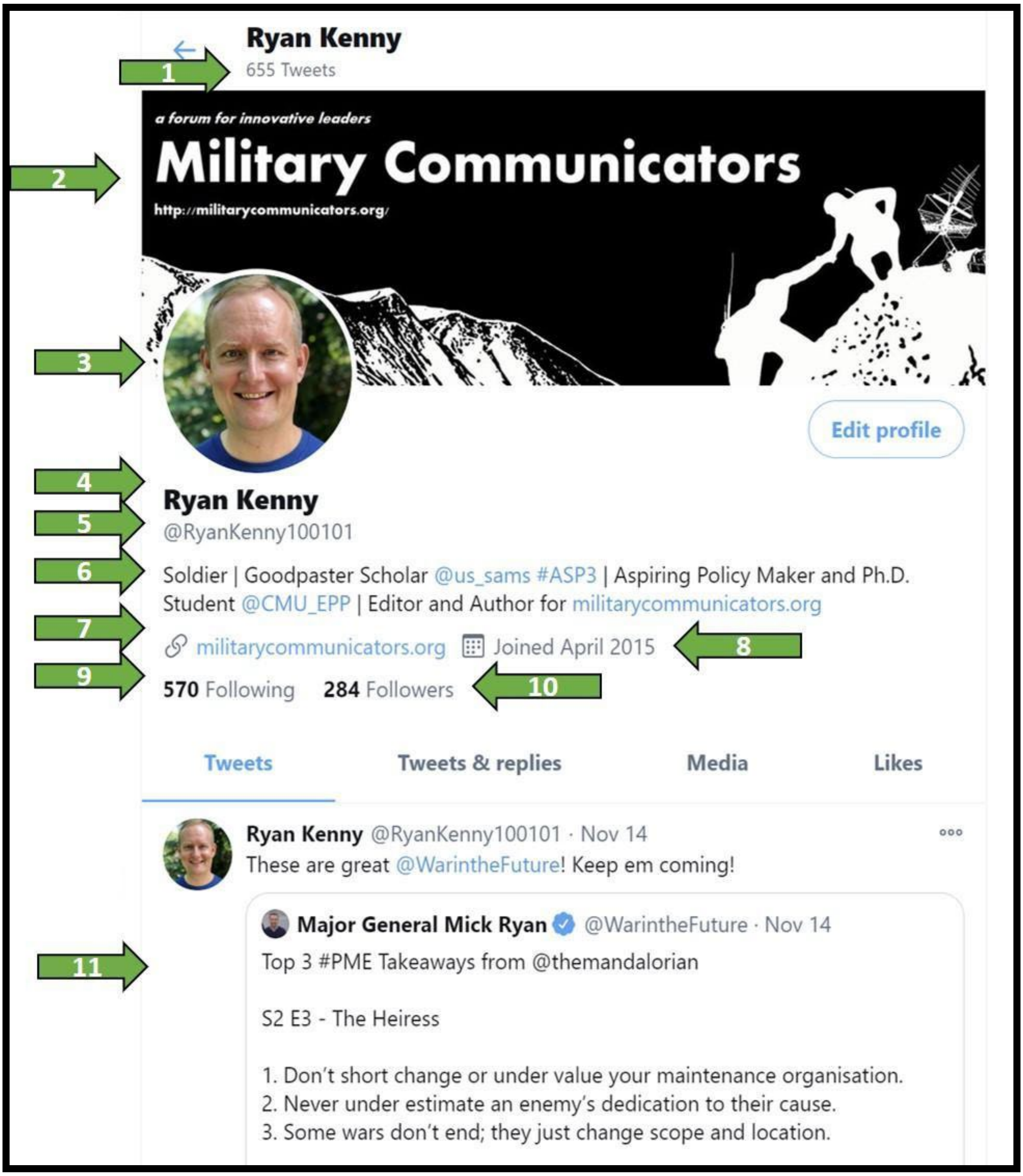
Data were collected in September 2020. Participants (N = 113) were recruited from U.S. Amazon Mechanical Turk (mTurk) and paid $15 for approximately 25 minutes of work. Mechanical Turk samples tend to be more varied than other convenience samples. These participants are not representative of the U.S. population (Crump, McDonnell, & Gureckis, 2013; Mason, & Suri, 2012) but perform similarly to other populations recruited for online research tasks (Loepp, & Kelly, 2020). Participation was limited to U.S. citizens and native English speakers. Informed consent was obtained. This research complied with the American Psychological Association Code of Ethics and was approved by the Carnegie Mellon University Institutional Review Board under protocol # IRB00000352.

Design

As shown in Figure 1, participants judged 52 Twitter persona profiles, each with 11 features (e.g., profile image, description, follower count). Although all personas were active from real-world accounts when their bot indicator scores were obtained (see above), participants were not explicitly told this. Twenty-five trials were “bot” personas; twenty-five were “human,” as defined by bot indicator scores above or below 50%. Participants made a binary judgment about whether each persona was a bot or a human, then gave the probability of that response being correct. The order of the stimuli was randomly determined for each participant. The two in-task attention check tasks were personas of public figures (Elizabeth Warren and Mike Pence), presented at random locations among the stimuli. Response time was collected and analyzed for both exclusion criteria and posthoc analyses. Full instructions appear in the Supplementary Materials (SM).

Figure 1

Example of Twitter Persona Profile



Ryan Kenny
655 Tweets

a forum for innovative leaders

Military Communicators

http://militarycommunicators.org/

Ryan Kenny
@RyanKenny100101

Soldier | Goodpaster Scholar @us_sams #ASP3 | Aspiring Policy Maker and Ph.D. Student @CMU_EPP | Editor and Author for militarycommunicators.org

militarycommunicators.org Joined April 2015

570 Following 284 Followers

Ryan Kenny @RyanKenny100101 · Nov 14
These are great @WarintheFuture! Keep em coming!

Major General Mick Ryan @WarintheFuture · Nov 14
Top 3 #PME Takeaways from @themandalorian

S2 E3 - The Heiress

1. Don't short change or under value your maintenance organisation.
2. Never under estimate an enemy's dedication to their cause.
3. Some wars don't end; they just change scope and location.

Note. All Twitter personas have features, as indicated by the arrows. Twitter provides some features: (1) the number of tweets a user has produced, (8) the date a user joined Twitter, (9) the number of other Twitter users the persona follows, and (10) the number of other users following the persona. The user provides others: (2) background image, (3) profile picture, (4) profile name, (5) the profile's Twitter label, (6) profile description, (7) linked personal pages, (11) and a pinned or the last tweet.

After completing the judgment task, participants completed a demographic survey and individual difference measures of (a) social media experience (Hou, 2017) (see Appendix B); (b) political views on a 5-point scale ranging from "liberal" to "conservative" (see Appendix C); and (c) cognitive reflection tendency, using the original three-item CRT (Frederick, 2005), (see Appendix D). A *political difference* score was created as the absolute difference between the participant's self-reported political views and each stimulus's political tone.

Analyses

Our planned analyses examined the contributions of task characteristics (bot indicator, stimulus presentation order), individual factors (task engagement, social media experience, cognitive reflection score, political difference), and their interactions to predict the sensitivity and criterion of participants' judgments of whether each stimulus was a bot or human persona. The post-hoc analysis examined relationships between participants' political views and their performance.

Traditional SDT analyses estimate sensitivity by calculating the standardized difference between each participant's 'hit' rate and 'false alarm' rate, then inferring the distributions for signal present and signal absent. This approach does not readily lend itself to examining how sensitivity and criterion are related to other variables. Therefore, we used a generalized linear mixed-effects probit regression to predict participants' probability of calling a persona a bot. This method allows unobserved heterogeneity in both the intercept (capturing the criterion) and slope (capturing sensitivity to bot indicator score), assuming a multivariate normal distribution (DeCarlo, 1998). Equations (1) and (2) show the planned and post-hoc models, following Farewell, Long, Tom, Yiu & Su (2017):

$$probit\{Pr(Yi,j = 1||Xi,j,Ui)\} = \theta Xi,j + Ui \tag{1}$$

$$\begin{aligned} &P('bot' \mid BI_i, TE_i, TO_i, SME_i, CRT_i, PV_i, PD_i) = \\ &\Phi\,(\beta_0 + \beta_1 BI_i + \beta_2 TE_i + \beta_3 TO_i + \beta_4 SME_i + \beta_5 CRT_i + \beta_6 PV_i + \\ &\beta_7 PD_i + \beta_8 TE_i * BI_i + \beta_9 TO_i * BI_i + \beta_{10} SME_i * BI_i + \beta_{11} CRT_i * BI_i + \\ &\beta_{12} PV_i * BI_i + \beta_{13} PD_i * BI_i + \beta_{14} PD_i * CRT_i * BI_i + \beta_{15} PV_i * CRT_i * \\ &BI_i + \beta_{16} PD_i * PV_i * BI_i + \beta_{17} PV_i * PD_i * CRT_i * BI_i) \end{aligned} \tag{2}$$

In Equation 1, $Y_{i,j}$ is the dependent variable observed for a participant *i*, for stimulus *j*, modeled using mixed probit regression with a random intercept and U_i for differences in participants' reactions to experimental conditions and predictor variables. The vector of predictor variables is denoted by $X_{i,j}$, with associated parameter vector Ѳ. Equation 2 shows how the probability of responding 'bot' is computed, applying the probit link function to the set of conditioned predictor variables. The coefficients in Equation 1 are presented in Table 1. Bot probabilities are calculated by applying a univariate Gaussian CDF to the linear predictor for each participant-persona combination. Random effects are modeled with a multivariate Gaussian distribution and estimated using the arm package in R (Gelman, Su, Yajima, Hill, Pittau, Kerman & Dorie, 2016).

The intercept in these models estimates the criterion for responding "bot" when all other regressors are set at their mean values (using normalized values with mean = 0). An intercept of 0 implies no preference for responding either "bot" or "human." A negative intercept indicates a tendency to judge personas as humans (i.e., more robust evidence is needed to say "bot" compared to a neutral criterion). A positive intercept indicates a tendency to judge personas as bots (i.e., weaker evidence is needed to say "bot"). A main effect reflects a criterion shift regardless of a stimulus's bot indicator score.

The bot indicator score is the probability that a stimulus is a bot, as determined by two machine learning algorithms. When the other regressors are set to zero, the average sensitivity in the sample is the coefficient on the bot indicator score (BI), corresponding to the change in the mean of the Gaussian distribution, as the BI score changes from 0 (definitely not a bot) to 1.0 (definitely a bot). Interactions with the BI score represent variations in sensitivity (d') among sample subgroups.

We used task order (TO) and task engagement (TE) as covariates to capture participant fatigue and concentration. Task order is the stimulus presentation order (from 1 to 52). Task engagement equaled the number correct on three attention checks (one after the practice trials and two during the experimental trials).

Table 1 presents the planned and post-hoc analyses results, with the latter adding respondents' political values (PV).

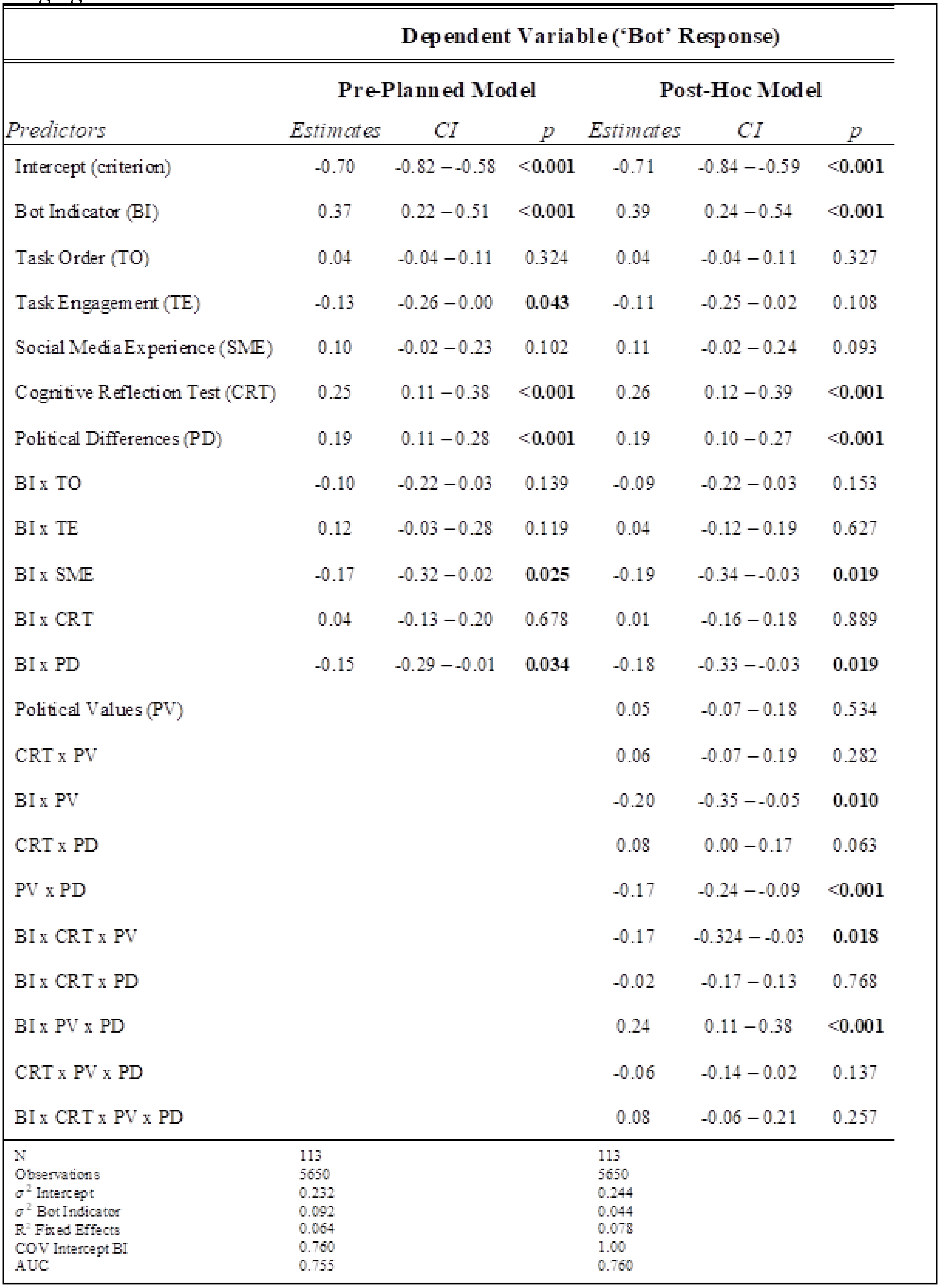
Results

Sample Demographics

The sample included 113 participants, 73 men and 40 women, whose ages ranged from 18 to 72 years old (mean = 36, median = 33). Eighty-seven reported being White, 5 Hispanic or Latino, 12 Black or African American, 1 Native American, 7 Asian or Pacific Islander, and 1 Other. Twenty-two reported a high school degree or equivalent, 73 a bachelor's degree, and 18 a master's degree. Ninety-four reported being fully employed, 7 employed part-time, 1 unemployed but looking, 2 retired, 8 were self-employed, and 1 unable to work. Sixty-eight reported being married, 5 divorced, 1 separated, and 39 never married. Annual incomes were roughly normally distributed, ranging from "less than 10K" to "over 150K," with the median between 50K-60K.

Table 1

General Linear Mixed Effects Probit Regression Models, Predicting the Probability of Judging a Persona to Be a Bot.



	Dependent Variable ('Bot' Response)					
	Pre-Planned Model			**Post-Hoc Model**		
Predictors	*Estimates*	*CI*	*p*	*Estimates*	*CI*	*p*
Intercept (criterion)	-0.70	-0.82 – -0.58	**<0.001**	-0.71	-0.84 – -0.59	**<0.001**
Bot Indicator (BI)	0.37	0.22 – 0.51	**<0.001**	0.39	0.24 – 0.54	**<0.001**
Task Order (TO)	0.04	-0.04 – 0.11	0.324	0.04	-0.04 – 0.11	0.327
Task Engagement (TE)	-0.13	-0.26 – 0.00	**0.043**	-0.11	-0.25 – 0.02	0.108
Social Media Experience (SME)	0.10	-0.02 – 0.23	0.102	0.11	-0.02 – 0.24	0.093
Cognitive Reflection Test (CRT)	0.25	0.11 – 0.38	**<0.001**	0.26	0.12 – 0.39	**<0.001**
Political Differences (PD)	0.19	0.11 – 0.28	**<0.001**	0.19	0.10 – 0.27	**<0.001**
BI x TO	-0.10	-0.22 – 0.03	0.139	-0.09	-0.22 – 0.03	0.153
BI x TE	0.12	-0.03 – 0.28	0.119	0.04	-0.12 – 0.19	0.627
BI x SME	-0.17	-0.32 – 0.02	**0.025**	-0.19	-0.34 – -0.03	**0.019**
BI x CRT	0.04	-0.13 – 0.20	0.678	0.01	-0.16 – 0.18	0.889
BI x PD	-0.15	-0.29 – -0.01	**0.034**	-0.18	-0.33 – -0.03	**0.019**
Political Values (PV)				0.05	-0.07 – 0.18	0.534
CRT x PV				0.06	-0.07 – 0.19	0.282
BI x PV				-0.20	-0.35 – -0.05	**0.010**
CRT x PD				0.08	0.00 – 0.17	0.063
PV x PD				-0.17	-0.24 – -0.09	**<0.001**
BI x CRT x PV				-0.17	-0.324 – -0.03	**0.018**
BI x CRT x PD				-0.02	-0.17 – 0.13	0.768
BI x PV x PD				0.24	0.11 – 0.38	**<0.001**
CRT x PV x PD				-0.06	-0.14 – 0.02	0.137
BI x CRT x PV x PD				0.08	-0.06 – 0.21	0.257
N	113			113		
Observations	5650			5650		
σ^2 Intercept	0.232			0.244		
σ^2 Bot Indicator	0.092			0.044		
R^2 Fixed Effects	0.064			0.078		
COV Intercept BI	0.760			1.00		
AUC	0.755			0.760		

Note. Table 1 shows results from pre-planned and post-hoc models predicting whether participants respond 'bot' = 1, vs. 'human' = 0, as a function of task characteristics (bot indicator, task order), individual factors (task engagement, social media experience, cognitive reflection test, political views), and political difference (between individual and stimulus). The coefficients are estimated with a general linear mixed-effects probit model. The dependent measure was each participant's judgment that a stimulus was a 'bot' or a 'human' persona. Bot indicator (BI) is the algorithm-derived probability of the stimulus being a bot. All other predictive variables were converted to *z*-scores for ease of interpretation. The analysis is based on N=113 participants. The intercept reflects the criterion for the average participant (with all other regressors at their mean), with higher values indicating a greater tendency to respond 'bot.' Positive interactions with the bot indicator score represent positive changes in participant sensitivity (d'). The AUC was computed with a BI threshold of 0.5.

Criterion, Bot Indicator, and Controls

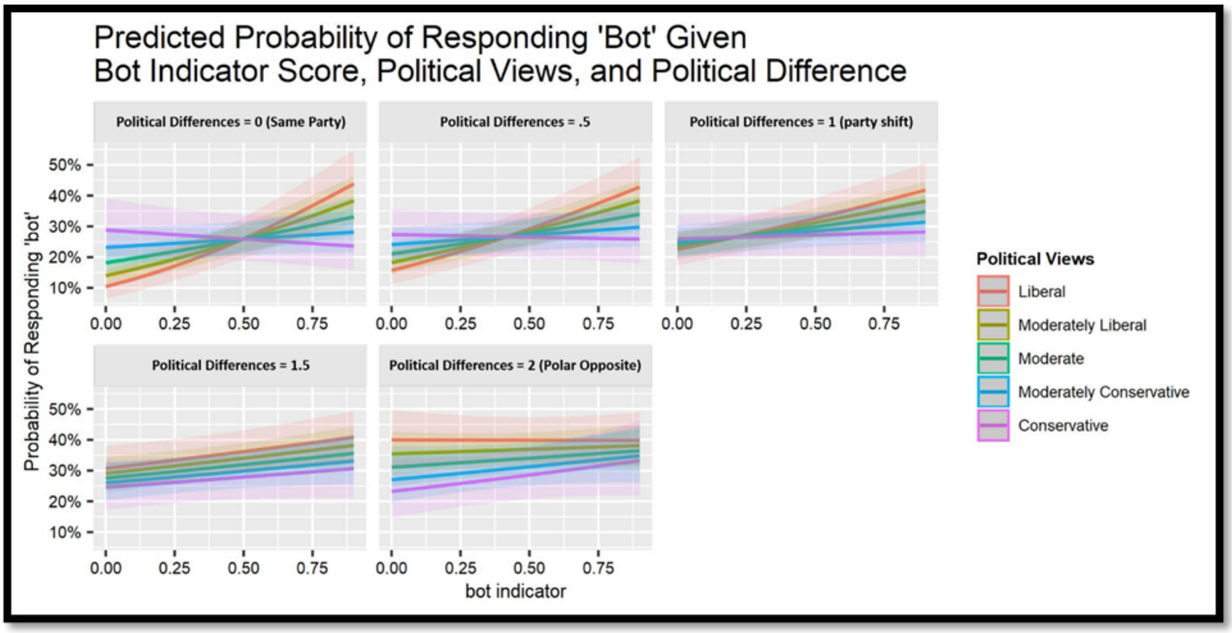
In both models, the intercepts were strongly negative (-0.70 and -0.71, respectively), indicating a tendency to respond 'human' when the other regressors were at their mean levels, even though half the stimuli were most likely bots. Both models found sensitivity (d'), as reflected in a positive coefficient for BI (0.37 and 0.39, respectively) when the other regressors were at their mean levels, meaning that responding 'bot' was more likely as a persona's bot indicator score increased ($p < 0.001$). Task order (TO) was unrelated to participants' probability of responding 'bot,' suggesting no sign of fatigue. For the planned model, the Task Engagement (TE) coefficient was –0.13 ($p = .043$)*,* indicating that participants were more likely to say human when more engaged in the task. None of these relationships depended on the likelihood of a persona being a bot (as seen in non-significant interactions with BI).

Political Values and Political Differences

As participants' political difference (PD) from the persona increased, they were more likely to judge it a 'bot,' consistent with myside bias, with a one standard deviation increase in PD shifting the intercept by 0.19. The post-hoc model adds participants' self-reported political values (PV) as the main effects and interactions. In the post hoc model, both liberal and conservative participants had a greater probability of responding 'bot' when viewing a persona of an opposing political view. However, liberals had a greater probability of responding 'bot' than conservatives.

Both models' interaction between bot indicator and political differences (BI x PD) is explained by adding political values in the post-hoc model. The significant three-way interaction (BI x PV x PD) (p < 0.001) reveals an asymmetric pattern of sensitivity related to participants' political views. Figure 2 shows the relationship between PV and BI, with PD divided into five levels. In the upper left, when judging personas with similar political views, liberals were very sensitive to the bot indicator score (red line), while conservatives were not (purple line). At the other extreme, when judging personas with opposite political views, liberals were insensitive to bot indicator scores, while conservatives had a modest sensitivity.

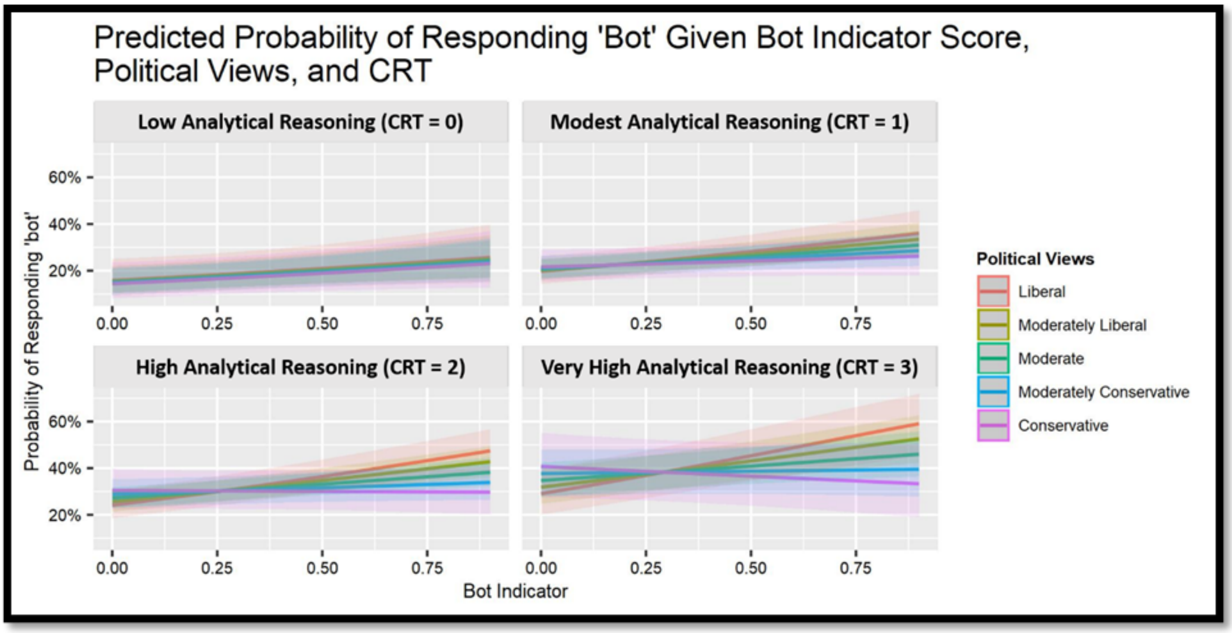
Figure 2



Political Values and Cognitive Reflection

In both models, participants with higher CRT scores were more likely to respond 'bot,' shifting the intercept by 0.26 in the final model for each standard deviation above the mean CRT score. The post-hoc model also reveals a significant (p = 0.018) three-way interaction between Political Values, Cognitive Reflection, and bot indicator (BI x CRT x PV). Figure 3 shows sensitivity to BI for participants with different PVs for the four possible CRT scores. Participants with the lowest CRT scores were largely insensitive to BI, whatever their politics. As the CRT score increased, so did sensitivity to BI for liberals but not for conservatives.

Figure 3



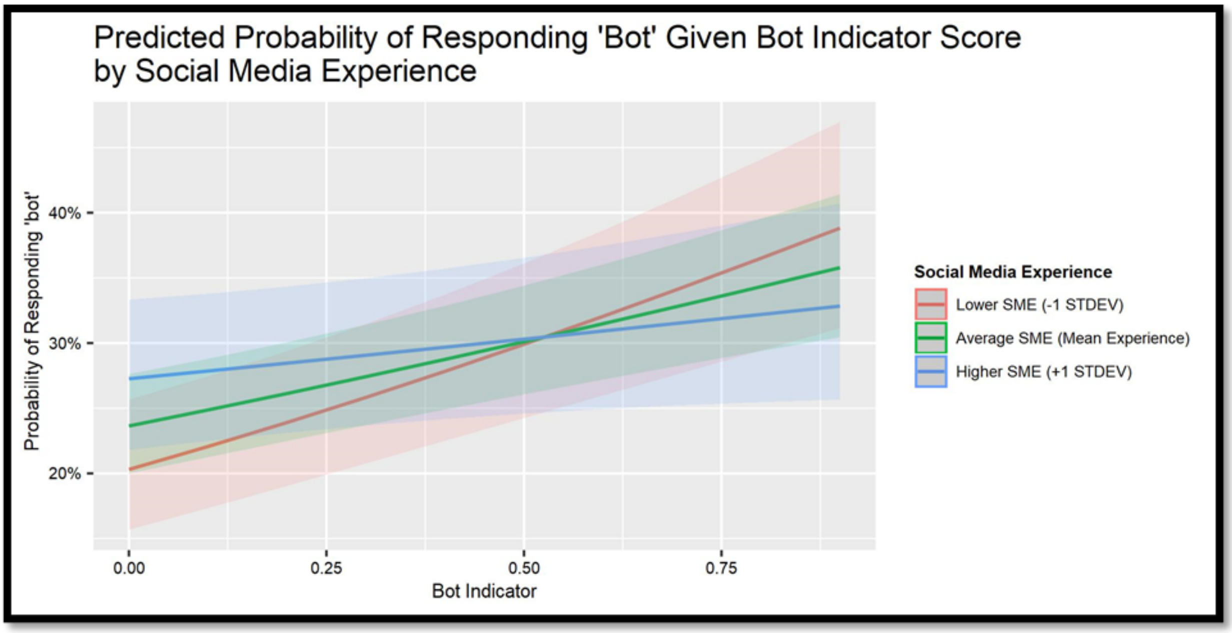
Social Media Experience

In both the planned and the post-hoc models, self-reported Social Media Experience (SME) was unrelated to the likelihood of calling personas 'bots.' In both models, there was a significant interaction with bot indicator (BI x SME) ($p < 0.025$ and 0.019 respectively), revealing a counter-intuitive finding. Figure 4 shows participants who reported SME 1 σ below and 1 σ above the sample mean. Those reporting higher SME were less sensitive to the bot indicator than were those reporting lower SME.

Confidence

Participants expressed moderate confidence in their bot/human judgments, M=81.2% (SD=13.2%), on the 50-100% scale. Overall, 52.5% of participants' judgments were accurate. Participants did not have particularly high sensitivity, as seen in the sensitivity findings. Thus, participants were overconfident in their abilities (28.7%) (Lichtenstein et al., 1982).

Figure 4



Discussion

This study assesses human performance in detecting bots among Twitter personas. Participants judged whether each of the 52 personas was produced by a human or a bot. The personas were chosen to represent a flat distribution of bot-indicator (BI) scores, reflecting their probability of being bots, as determined by agreement between two machine learning algorithms.

Planned Analyses

Sensitivity to Bot Indicator scores (BI). We found that participants were sensitive to the differences between bot and human personas, as reflected in a modest positive correlation between BI scores and participants' probability of saying 'bot.' That sensitivity varied by other individual characteristics (as described below). It was unrelated to our measures of Task Order (TO), meant to assess the effects of fatigue, Task Engagement (TE), meant to assess attention, either directly or in interaction with BI (BI x TO, BI x TE).

Criterion for Responding 'Bot'. Although bot and human personas were equally likely, and participants were cautioned to be on guard for bots, participants judged a majority to be humans. If they treated the two categories as equally likely, the intercept of the prediction model would be 0.0. However, we observed an intercept of -0.71. Holding all other predictor variables constant, including the bot indicator score, equates to a 76%

probability of responding ‘human.’ If participants assumed that humans and bots were equally likely, they were more averse to mistaking a human for a bot than vice versa. Alternatively, they may have had strong prior beliefs that most personas are human in the study and perhaps the world.

Self-reported social media experience (SME). Studies typically find that experts outperform novices in discrimination tasks (Allen, Mcgeorge, Pearson & Milne, 2004; Bond, 2008; Spengler, White, Ægisdóttir, Maugherman, Anderson, Cook & Rush, 2009; Cañal-Bruland & Schmidt, 2009). However, we found the opposite: participants who reported greater social media experience were less sensitive (Figure 4). One possible explanation is that such experience does not confer expertise (Ericsson, 2018), joining the handful of other studies in which novices outperform experts (Bisseret, 1981; Rikers, Schmidt & Boshuizen, 2000; Witteman & Tollenaar, 2012). Frequent users may have convinced themselves that they can tell a bot from a human without clear feedback to prove them wrong.

Cognitive Reflection Test (CRT). The Cognitive Reflection Test is meant to assess individuals' willingness and ability to resist false lures and find correct answers to narrative problems. CRT scores were, however, not related to sensitivity, except in their interaction with political differences (Figure 3).

Political Differences (PD). We calculated the absolute difference between the participant's self-reported political value (PV) and the persona’s. Participants were more sensitive to the properties captured by the bot indicator score when they agreed a persona’s political tone (Figure 2; PD=0) than when they disagreed (Figure 2: PD=2). When combined with the criterion shift toward a more liberal tendency to respond ‘bot’ when viewing personas of opposing political views, we interpreted this result as reflecting myside bias, the tendency to look harder at contrary evidence.

Post Hoc Analyses

Post-hoc analyses revealed two statistically significant three-way interactions. One (BI x PV x PD) found that liberals were more sensitive to bot indicator (BI) than conservatives when judging politically similar personas. The second (BI x CRT x PV) found that the judgments of participants with low CRT scores were unrelated to their political viewpoint (PV); however, for participants with high CRT scores, liberals were sensitive to bot indicator (BI), whereas conservatives were not.

We conducted this study in Fall 2020, at a time of high political distrust (Gramlich, 2020; Iyengar, Lelkes, Levendusky, Malhotra, & Westwood, 2019). These results are consistent with that distrust, revealing themselves somewhat differently with liberals than conservatives, suggesting either differences in reasoning styles (Deppe, Gonzalez, Neiman, Pahlke, Smith, & Hibbing, 2015) or political discourse at a time when some liberals accused conservatives of spreading misinformation during the 2018 and 2020 elections (Lee & Hosam, 2020). Liberal participants may have been predisposed to view conservative personas as social bots, whereas conservatives, defensive about the charge, may have been reluctant to label fellow conservatives as bots.

Previous studies have found conflicting results regarding the relationship between cognitive skills and the ability to detect misleading information. Pennycook and colleagues have found that people with higher CRT scores have more ability (Pennycook & Rand, 2019; Bronstein, Pennycook, Bear, Rand & Cannon, 2019; Ross, Rand & Pennycook, 2021). Other studies have found that people employ their cognitive skills to support their ideological views and biases (Drummond & Fischhoff, 2017; Haidt, 2012; Stanovich & West, 2007; Strickland, Taber & Lodge, 2011). The present results are consistent with the latter findings amongst liberals, as reflected in participants with higher CRT scores being more likely to treat personas of opposing political views as bots, but not for conservatives.

Application to Bot Detection

These findings highlight the importance and challenge of designing interventions to improve social bot detection. Social media users spend much of their online time in echo chambers with like-minded individuals, some of whom may be bots (Choi, Chun, Oh, & Han, 2020; Sasahara, Chen, Peng, Ciampaglia, Flammini & Menczer, 2019). If, as we found, detecting social bots is harder with ingroups than with outgroups, then heavy social media users may be particularly vulnerable to being duped by bots that look like people they trust. That vulnerability may grow with time spent in their "bubble," as seen in the poorer performance of participants reporting greater social media experience. They might be especially advised to beware of complacency and seeming friends.

Limitations and Future Work

Our conclusions depend upon the accuracy of the normative bot indicator scores provided by the two machine learning systems, Botometer and Bot-hunter. Like other machine learning models, they may have been trained on unrepresentative and mislabeled training sets, with unknown effects on our results. As indirect evidence of their validity, we examined the personas used here in March 2021, approximately one year after their Twitter profiles were initially assessed (see Figures 5A and 6A, Appendix A). Among the 13 personas with bot indicator scores over 75%, seven (54%) had been suspended by Twitter, one no longer existed, and three of the five remaining had lost an average of 36% of their followers. For the 24 personas with bot indicator scores between 25% and 75%, 5 (16%) had been suspended; only 4 of the other 20 had lost more than 36% of followers. Among personas with bot indicator scores less than 25%, there were no suspended accounts and no significant change in their number of followers. Although accounts can be suspended and lose followers for reasons other than being bots, and bots can go undetected, these observations add credibility to the bot indicator scores used here.

A second potential limitation is our experimental task. As with other simulated experiences (Wald, Khoshgoftaar, Napolitano, & Sumner, 2013; Aiello, Deplano, Schifanella, & Ruffo, 2012), the validity of our task depends on how well it evoked real-world behavior. We used actual personas and set a pace akin to the rapid evaluation of everyday Twitter use. However, we did not provide access to the persona profile pages that suspicious users might examine; hence might have underestimated their abilities. We may have had a higher proportion of social bots than that observed in everyday life, contributing to the tendency to identify them as human (Varol, Ferrara, Davis, Menczer & Flammini, 2017). Further research would be needed to examine these possibilities.

That research might also try to understand why self-related social media experience was related to poorer performance. If that result proves robust, social media platforms may use automated systems to identify likely social bots, providing feedback that is currently unavailable. Platforms might also experiment with encouraging people to be less trusting of the authenticity of bots that seem to share their political views.

Conclusion

Social bot developers are becoming increasingly sophisticated at mimicking human personas and manipulating users' commercial and political behavior. Our results suggest that, even with today's social bots, people need help, especially with bots that prey on the false sense of security that comes with social media experience and engaging bots that express their political orientation. That help might come in the form of warnings about myside bias or bot indicator scores. Evaluating such interventions is an urgent question for protecting people from social media manipulation.

Key Points

- We evaluated performance in distinguishing Twitter personas produced by humans and social bots.
- We found relatively low sensitivity, as reflected in correlations between participants' judgments and bot indicators scores produced by two machine learning algorithms.
- We found a greater aversion to mistaking a human for a bot than mistaking a bot for a human.
- We found evidence of myside bias, with individuals being less critical of bots that shared their political values.
- We found poorer performance among participants who reported more social media experience.

Chapter 3. Aiding Social Bot Detection

In Study 1 (Kenny et al., 2022), participants examined Twitter profiles with varying bot signals and judged them as humans or social bots. On average, participants had modest sensitivity (d') and a criterion that favored responding 'human.' Consistent with myside bias, that criterion shifted toward greater acceptance of personas as humans when participants shared the persona's political views. Sensitivity to social bots increased with greater task engagement, analytical reasoning ability (CRT scores), and less social media experience. Political conservatives displayed less sensitivity to social bots that shared their political views. These findings reveal the significant risk of social bot deception and a pressing need to safeguard people from being duped by social bots.

Computer scientists have sought to improve social bot detection using Artificial Intelligence (AI) systems for over a decade (Cresci, 2020; Davis, Varol, Ferrara, Flammini & Menczer, 2016; Beskow & Carley, 2018). These AI systems are automatic algorithms that adjust their parameters and predictions in response to changes in data (Samuel, 1959; Mitchell, 1997). In the context of social bot detection, these algorithms aim to minimize misses and false alarms while maximizing hits and correct rejections. After training, they can aid in pattern recognition, rule induction, and event prediction.

Many social scientists have touted the benefits of algorithms to overcome human judgment and decision-making limitations (e.g., Dawes, 1979; Meehl, 1954). AI systems have been created to aid decision-making in a wide range of contexts, including criminal justice (Kluttz, & Mulligan, 2019), investing (Andriosopoulos, Doumpos, Pardalos, & Zopounidis, 2019), and healthcare (Sutton, Pincock, Baumgart, Sadowski, Fedorak, & Kroeker, 2020). However, even when these systems demonstrably outperform humans, they have not always been readily accepted by users.

There are currently several algorithmic tools for detecting Twitter social bots. Some of these tools require users to visit a third-party website, enter a username, and interpret a set of findings (Botometer, Bot Sentinel). Others offer overlays and inline bot indicator signals within the Twitter platform. For instance, in Fall 2021, Twitter began testing bot indicator icons near profile user names (Twitter, 2022). NortonLifeLock has developed an add-on showing both a bot indicator icon and a probability score derived from their proprietary social bot detection algorithm (Kats, 2022). If social media users will and can use them, these inline bot detection aids could offer easy, effective means to protect themselves from deception by fake personas and social bots. However, research has found that, even when afforded access to valid algorithmic predictions, many people choose not to use them (Dietvorst, Simmons & Massey, 2015; Burton, Stein & Jensen, 2019; Diab, Pui, Yankelevich, & Highhouse, 2011).

This study asks how well people will use an algorithmic aid to social bot detection and which individual characteristics influence their performance. We also explore their intended social media behavior after completing the social bot detection task and their willingness to pay for social bot detection aid. Our approach adapts signal detection theory (SDT) to a simulated version of the rapid-fire decisions common to processing Twitter profiles (Kenny et al., 2022).

Early studies of users' willingness to accept aid from technologies focused on their opinions of technological features, especially perceived usefulness and ease of use (i.e., Davis, 1989). Recent research on user acceptance of advisor systems has centered on users' trust in the algorithm's tangibility, transparency, reliability, and task characteristics, such as technicality, data analysis and social intelligence requirements (e.g., Glikson & Woolley, 2020, Ghazizadeh, Lee, & Boyle, 2012; Hoff & Bashir, 2015; Lee & See, 2004; Pavlou, 2003). These studies have found that willingness to accept advice from AI systems depends primarily on users' (a) perceptions of system credibility, (b) perceptions of system applicability to specific tasks, and (c) ability to process systems' sometimes uncertain signals (Sembroski, Fraune & Šabanović, 2017).

Regarding (a), we experimentally vary the properties of the system. Regarding (b), we examine myside bias, a well-established psychological process whereby people are less critical when evaluating evidence consistent with their prior beliefs. Regarding (c), we consider individual differences in participants' analytical reasoning and social media experience.

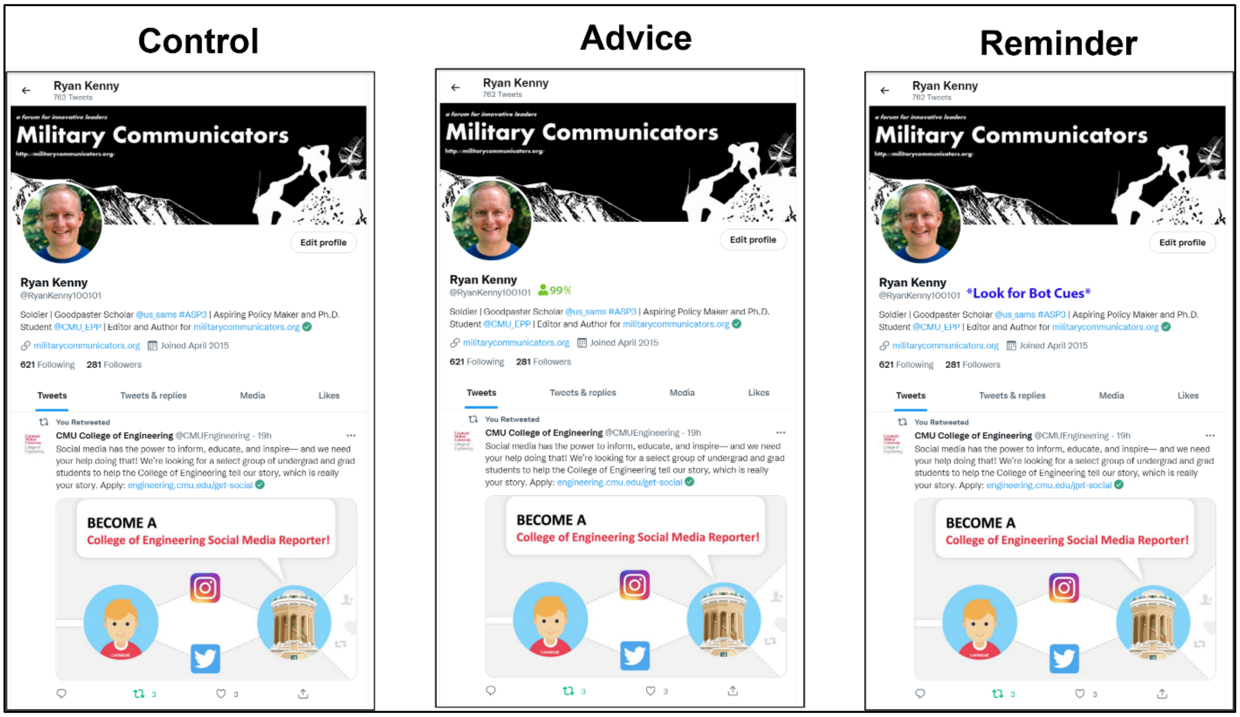
We also consider two procedural features that might affect performance: fatigue, as reflected in the task order, and engagement, as reflected in attention checks. We expect performance to decline the longer participants have worked and the less they are engaged. After introducing the task and our dependent measures of SDT performance, we present the variables in each class and the predicted relationships. Unless described as exploratory, all hypotheses were preregistered.

Task

Within a simulated Twitter setting, participants judged a series of Twitter persona profiles as either human or bot under three conditions, without aid (Control), with aid (Aid), or with a reminder to look for bot cues (Reminder); see Figure 1. After judging each profile, participants rated their confidence in their choice and then said whether they would share (retweet) a message from that persona if they agreed with its content. After completing these tasks, participants stated their willingness to pay for such aid. They then completed demographic questions, a survey of social media experience, and a test of analytical reasoning ability.

Figure 1

Twitter Persona Profiles by Test Condition



Ryan Kenny
762 Tweets

a forum for innovative leaders
Military Communicators
http://militarycommunicators.org/

Ryan Kenny
@RyanKenny100101

Soldier | Goodpaster Scholar @us_sams #ASP3 | Aspiring Policy Maker and Ph.D. Student @CMU_EPP | Editor and Author for militarycommunicators.org

militarycommunicators.org Joined April 2015

621 Following 281 Followers

You Retweeted

CMU College of Engineering @CMUEngineering · 19h

Social media has the power to inform, educate, and inspire— and we need your help doing that! We're looking for a select group of undergrad and grad students to help the College of Engineering tell our story, which is really your story. Apply: engineering.cmu.edu/get-social

BECOME A
College of Engineering Social Media Reporter!

Ryan Kenny
762 Tweets

a forum for innovative leaders
Military Communicators
http://militarycommunicators.org/

Ryan Kenny
@RyanKenny100101 99%

Soldier | Goodpaster Scholar @us_sams #ASP3 | Aspiring Policy Maker and Ph.D. Student @CMU_EPP | Editor and Author for militarycommunicators.org

militarycommunicators.org Joined April 2015

621 Following 281 Followers

You Retweeted

CMU College of Engineering @CMUEngineering · 19h

Social media has the power to inform, educate, and inspire— and we need your help doing that! We're looking for a select group of undergrad and grad students to help the College of Engineering tell our story, which is really your story. Apply: engineering.cmu.edu/get-social

BECOME A
College of Engineering Social Media Reporter!

Ryan Kenny
762 Tweets

a forum for innovative leaders
Military Communicators
http://militarycommunicators.org/

Ryan Kenny
@RyanKenny100101 *Look for Bot Cues*

Soldier | Goodpaster Scholar @us_sams #ASP3 | Aspiring Policy Maker and Ph.D. Student @CMU_EPP | Editor and Author for militarycommunicators.org

militarycommunicators.org Joined April 2015

621 Following 281 Followers

You Retweeted

CMU College of Engineering @CMUEngineering · 19h

Social media has the power to inform, educate, and inspire— and we need your help doing that! We're looking for a select group of undergrad and grad students to help the College of Engineering tell our story, which is really your story. Apply: engineering.cmu.edu/get-social

BECOME A
College of Engineering Social Media Reporter!

Note. The bot indicator score in the aid condition and the reminder cue were placed in the same location to match expected visual search patterns.

Stimulus Selection

We asked participants to examine Twitter persona profiles, as tweets alone reveal little about a persona. These profiles are the natural place to look when users are concerned about a persona's legitimacy. We selected currently active Twitter personas. To estimate a persona's probability of being a social bot, we used three machine-learning social bot detection systems for Twitter: Bot-hunter (Beskow, & Carley, 2018), Botometer (Davis, Varol, Ferrara, Flammini & Menczer, 2016), and Bot Sight (Kats, 2022), (see Appendix A). Each produces a probability of a persona being a social bot. They were trained and developed independently.

We relied upon Bot Hunter's Tier 1 model (see Figures 1A and 2A, Appendix A), which uses features accessible to average users as the basis of our bot indicator score. We then sought corresponding ratings from Botometer and Bot Sight assessments, accepting personas where the probability scores for the three algorithms concurred. On average, Bot Hunter and Botometer scores had a 0.926 correlation, and Bot Hunter and Botometer scores had a 0.88 correlation.

We created a suite of stimuli whose bot indicator scores were roughly uniformly distributed from very low (1%) to very high (99%). Additionally, we selected personas so that an even number of liberal, moderate, and conservative personas were distributed above and below the 50% threshold. (Study 1 describes our procedure for characterizing political identity.)

Dependent Measures

Sensitivity. Participants' *sensitivity* is the correspondence between their categorical judgments (bot, human) and the bot indicator score. More sensitive participants' judgments should change more quickly as bot indicator scores change.

Criterion. Participants' *criterion* reflects their tendency to treat ambiguous personas as bots or humans. Participants with lower thresholds for responding 'bot' have a higher hit rate and more false alarms, reflecting a desire not to miss any bots. Participants with higher thresholds have a higher rate of correct rejections and misses, reflecting a desire not to treat a human as a bot.

System Properties: Experimental Manipulations

We compared the performance of participants randomly assigned to one of three groups, all of whom judged the same 60 Twitter persona in stimuli whose creation is described below. *Control* group participants judged the personas in Twitter profiles per se. *Aid* group participants judged the stimuli with a bot indicator score placed near the persona's name. *Reminder* group participants judged the stimuli with a standard reminder to look for bot cues. We had the following preregistered hypotheses regarding these differences in the following subsections.

(H1 – d') Based on Study 1 findings, as the strength of the bot signal increases, as estimated by the bot indicator score, the probability of participants responding 'bot' will increase in all three conditions.

(H2 – d') Aid group participants will use the bot indicator score effectively, increasing their sensitivity compared to the control and reminder groups.

(H3 – d'; exploratory) Reliance on the bot indicator score will increase as it approaches greater certainty (0% or 100%), consistent with the probability weighting function in prospect theory (Tversky, & Kahneman, 1992). Conversely, there will be less sensitivity as the bot indicator score approaches 50%. Appendix I provides our findings.

(H4 – d'; exploratory) Reminder participants will perform better than control group participants, reflecting the reminder’s global signal regarding the risk. They will perform less well than the aid group due to lacking stimulus-specific information and habituating to the unchanging message.

(H5 - criterion) Participants in Study 1 had an aversion to misidentifying humans as bots, responding "human" more often than "bot," despite the equal number of each in the stimulus set. We expect that pattern here as well.

(H6 – criterion) Participants in the *aid* and *reminder* conditions will have lower criteria when compared with the control condition, as both interventions caution users against uncritically accepting bots as humans.

Perceptions of Individual Stimuli: Myside Bias

Myside bias entails examining information less critically if its source or content supports one's views (Stanovich, West & Toplak, 2013; Stanovich, & West, 2008; Toplak, & Stanovich, 2003; Drummond & Fischhoff, 2019; Kahan, Landrum, Carpenter, Helft, & Hall Jamieson, 2017; Taber, & Lodge, 2006; Taber, Cann, & Kucsova, 2009). Study 1 found evidence of myside bias in participants’ lowering their criterion for personas that disagreed with their political beliefs, more readily calling them bots, and stronger bias for individuals with stronger political views (Kenny et al., 2022). Exploratory analyses found different patterns of myside bias for liberals and conservatives differently. Conservatives considering conservatives had decreased relative sensitivity. Liberals, conversely, had higher sensitivity while judging liberals. However, their lower threshold when judging conservatives meant that they also had lower sensitivity. Thus, by assessing sensitivity and criterion separately, the SDT paradigm may shed additional light on the processes involved in political polarization, which play out on Twitter and elsewhere.

(H7 - criterion) (a) We will observe myside bias. (b) It will be stronger for liberals than conservatives, replicating the exploratory analysis pattern observed in Study 1.

(H8 – d'). We expect to replicate the results from Study 1, such that conservatives' myside bias entails not just a shift in criterion but also greater sensitivity when judging liberals than conservatives.

As in Study 1, we expect no overall relationship between PV and sensitivity and criterion, with differences emerging only with PD.

Information Processing Ability: Individual Differences

Analytical Reasoning Ability. Higher scores on various forms of the Cognitive Reflection Test (CRT) have been associated with better performance on various discrimination tasks (Campitelli & Labollita, 2010; Bar-Hillel, Noah & Frederick, 2019; Toplak, West & Stanovich, 2011). Study 1 replicated that general result with the bot detection task. We expected to find the same effect here and replicated an unexpected correlation between CRT and criterion.

(H9) (a) Participants with higher CRT scores will have (a) greater sensitivity and (b) a lower criterion, reflecting greater willingness to respond 'bot.' We expect similar results in all conditions.

Exploratory analyses in Study 1 found that the relationship with CRT was stronger for politically liberal respondents than for politically conservative ones. As mentioned, we also found less myside bias in the decreased sensitivity in liberals. Both results appear consistent with political psychology claims that core values of conservatism define how individuals manage uncertainty and threats (Jost, Glaser, Kruglanski, & Sulloway, 2003, Jost, 2017; Kahan, 2013; Kahan, Peters, Dawson, & Slovic, 2013). Though these causal mechanisms remain unclear, to the extent that political views shape reasoning about ambiguous threats, we expect to see the same pattern here. We pool results across conditions to increase statistical power for a weak effect in Study 1 and have no reason to expect the intervention to affect it.

(H10a – d') The correlation between CRT scores and sensitivity will be higher for political liberals than political conservatives.

(H10b - criterion) The correlation between CRT scores and criterion will be higher for politically liberals judging conservative personas than politically conservatives participants judging liberal personas.

Social Media Experience. Participants with more social media experience would be expected to have greater sensitivity if that experience has provided useful feedback (Langley, 1985; Pressley, Borkowski, & Schneider, 1987; Ohlsson, 1996; Ericsson, Krampe & Tesch-Römer, 1993). However, Study 1 found that participants who reported greater social media experience were *less* sensitive to stimulus personas' bot indicator score. We speculated that social media provide such poor feedback that frequent use may induce an illusory feeling of mastery without any actual increase in discrimination ability. We expected that result to repeat here in the control and reminder conditions. We expected sensitivity to increase with social media experience in the aid condition, as experienced users will learn more from the implicit feedback provided by how well bot indicator scores match their intuitive judgments.

(H11) Self-reported social media experience will be associated with reduced sensitivity in the control and reminder conditions but greater sensitivity in the aid condition.

Controls. We include two control variables in our prediction models: (a) *Stimulus Presentation Order (SPO)* to see if fatigue reduces sensitivity (Parasuraman & Davies, 1977; Warm, Parasuraman, & Matthews, 2008); and (b) *Task Engagement (TE)*, to see if more engaged participants perform better, as seen in Study 1 and others (e.g., Matthews, Warm, Reinerman, Langheim & Saxby, 2010; Downs, Holbrook, Sheng, & Cranor 2010; Dewitt, Fischhoff, Davis & Broomell, 2015). We do not expect the experimental manipulations to affect these relationships.

(H12a). Participants' sensitivity will decline with task order.

(H12b) Participants who answer more attention checks correctly will demonstrate greater sensitivity.

(H12c) We expected no correlation between either control variable and participants' decision criteria.

Methods

Sample

We collected data in January 2022. Participants (N = 924) were recruited using Prolific (Palan, & Schitter, 2018) and paid $8 for approximately 30 minutes of work. Evidence suggests that Prolific participants perform as well, if not better than Mturk workers and university subject pools (Peer et al., 2021, Peer et al., 2017). Eyal et al. (2021) compared data quality for several online behavioral research platforms, including Mturk, CloudResearch, and Prolific, and found that only Prolific provided high data quality on all their measures.

Participation was limited to US citizens and native English speakers. Informed consent was obtained. The research followed the American Psychological Association Code of Ethics and was approved by the Carnegie Mellon University Institutional Review Board under protocol # IRB00000472.

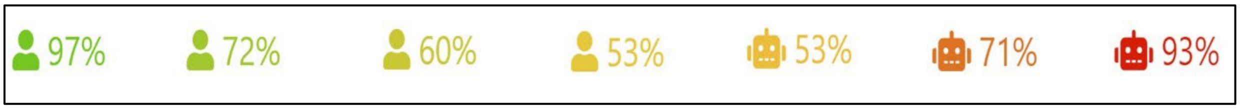
Design

Participants judged 60 Twitter personas like those in Figure 1, each characterized by 11 features (e.g., profile image, description, follower count). Participants were randomly assigned to view the personas with no aid (Control), a reminder (Reminder), or the bot indicator score aid (Aid). Aid group participants received each persona’s bot indicator score. Reminder group participants received the caution seen in Figure 1, similarly placed in each stimulus.

Figure 2 depicts the icons used in the aid condition. These were adapted from Bot Sight’s beta social bot indicator add-in (Kats, 2022). Personas with bot probabilities greater than 50% had a robot icon placed next to the persona's name, alongside the accompanying bot indicator score. Personas with bot probabilities below 50% had a human icon appear next to the persona's name with the bot indicator score’s complement (e.g., a bot indicator score of 10% was transformed to a human indicator score of 90%). The color scheme below was used to enhance the message.

Figure 2

Bot Indicator Aid Icons



Note: The bot indicator icons and coloring are based on NortonLifeLock Bot Sight: https://www.nortonlifelock.com/blogs/norton-labs/botsight-tool-detect-twitter-bots

Stimulus selection. The bot indicator scores were uniformly distributed between 1% and 99%, as determined by Bot Hunter and validated by Bot Sight and Botometer. Appendix A describes the selection process in greater detail. The political tone of the profiles was divided evenly for both bot and human personas, with one-third each being conservative, independent, and liberal, and with equal numbers above and below the 50% threshold. Each participant received 30 'bot' personas and 30 'human' personas, whose order was randomly determined. Three additional interspersed trials presented personas of public figures (Elizabeth Warren, Mitch McConnell, and Kim Kardashian) as attention checks.

Task. In the real world, a user who wished to investigate whether a persona was a human or a bot would start by examining the persona's profile page. We sought to simulate this setting. In each trial, participants examined a public Twitter persona profile and indicated whether they believed that persona was created by a 'bot' or a human. They then rated their confidence in this choice, using a slide bar on a scale anchored at 50% (completely uncertain) and 100% (certain). Finally, they indicated whether they would 'share' a tweet from the persona by retweeting it if they agreed with its content. We collected response times for these answers as exploratory measures of effort and possible future attention checks.

After completing the trials, participants completed a demographic survey and individual difference measures of (a) social media experience (Hou, 2017), (see Appendix B), (b) political views on a 7-point scale ranging from "strongly liberal" to "strongly conservative," (see Appendix C); (c) analytical reasoning ability, using combined scores from a modified three-item Cognitive Reflection Test (CRT) (Frederick, 2005) and four-item CRT developed by Thomson and Oppenheimer (2016), (see Appendix D). Both versions are reliable and have correlated scores (Pennycook & Rand, 2019). A political difference score (PD) was calculated by taking the absolute difference between participants' self-reported political views and each stimulus's political tone (see SM for rating procedure). Participants then completed measures of (d) concern over social bots, see Appendix E, and (e) willingness to pay (WTP), see Appendix F, for a social bot detection service.

Analysis Plan

Signal Detection Task

Traditional SDT analyses calculate the difference between standardized distributions of each participant's 'hit' rate and 'false alarm' rate to infer the distributions for signal present and signal absent and, thereby, estimate sensitivity. This approach works well when averaging sensitivity across trials. However, it does not readily lend itself to examining stimulus effects (e.g., political tone, presentation order). We used a generalized linear mixed-effects probit regression to assess trial-level effects, predicting participants' probability of calling each persona a bot. This approach uses the unobserved heterogeneity in both the intercept and the slope to capture the criterion and sensitivity to the bot indicator, assuming a multivariate normal distribution (DeCarlo, 1998).

All models employed a probit link function. To predict the probability of responding 'bot,' z-scores for each observation are converted to possibilities by taking the inverse of Phi. Random effects for each model coefficient are estimated with a multivariate Gaussian distribution, using the arm package in R (Gelman, Su, Yajima, Hill, Pittau, Kerman & Dorie, 2016).

Our planned (preregistered) analyses examined the contributions of the experimental conditions (control, aid, reminder), stimuli attributes (bot indicator score, political difference), and individual characteristics (social media experience, CRT score, and political view) in determining the probability of responding 'bot.' Stimulus presentation order and participant task engagement were included as control variables.

In these models, when all other regressors are set at their mean values (using normalized values with mean = 0), the intercept estimates the criterion for responding "bot." An intercept of 0 indicates that a participant is equally likely to say "bot" or "human." A negative intercept suggests a more lenient response criterion (i.e., requiring stronger evidence to call a persona a "bot"). A positive intercept indicates a more stringent criterion (i.e., requiring stronger evidence to say "human").

We treated the bot indicator score as the probability that a stimulus is a bot. When the other regressors are set to zero, the average sensitivity within a condition is the coefficient on the bot indicator score (BI). As the BI score changes from 0 (definitely not a bot) to 1.0 (definitely a bot), the value of the coefficient corresponds to the change in the mean of the Gaussian distribution. The interactions of other predictor variables with the BI score represent variations in sensitivity (d') among sample subgroups.

The regressions will introduce variables in five cumulative blocks, predicting whether participants respond to 'bot' = 1 or 'human' = 0. Model 1's predictors are test conditions (control, aid, reminder) and controls (task order, task engagement). Model 2 adds the individual characteristic of social media experience. Model 3 adds the individual characteristic of cognitive reflection. Model 4 adds the individual characteristic of political views. Model 5 adds the political difference between individual and stimulus.

Bot indicator (BI) is the algorithm-derived probability of the stimulus being a bot in each model. The intercept reflects the criterion for the average participant in the control condition (with all other regressors at their mean), with higher values indicating a greater tendency to respond 'bot.' Positive interactions with the bot indicator score reflect increased sensitivity (d') for a given subgroup of participants.

Results

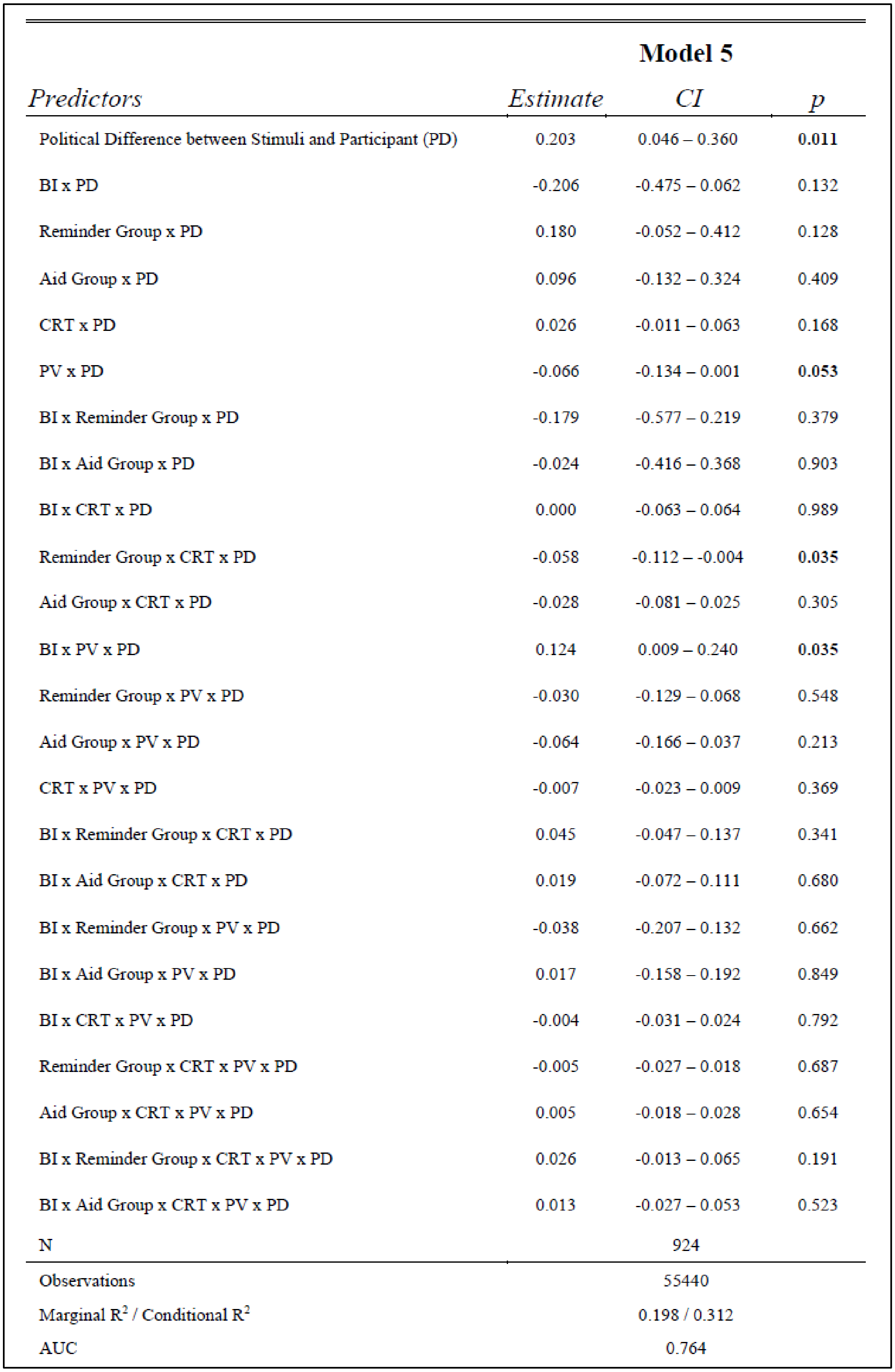
Tables 1a and 1b present the planned (preregistered) SDT analyses.

Table 1*a*

General Linear Mixed Effects Probit Regression Models, Predicting the Probability of Judging a Persona to Be a Bot.

Dependent Variable = ('Bot' Response)															
	Model 1			Model 2			Model 3			Model 4			Model 5		
Predictors	*Estimate*	*CI*	*p*	*Estimate*	*CI*	*p*	*Estimate*	*CI*	*p*	*Estimate*	*CI*	*p*	*Estimate*	*CI*	*p*
(Intercept) (Control Group Criterion)	0.094	-0.166 – 0.354	0.478	0.120	-0.195 – 0.435	0.455	0.054	-0.284 – 0.392	0.754	-0.006	-0.351 – 0.338	0.971	-0.186	-0.565 – 0.192	0.335
Bot Indicator (BI)	-0.501	-0.943 – -0.058	**0.027**	-0.614	-1.151 – -0.077	**0.025**	-0.579	-1.151 – -0.007	**0.047**	-0.437	-1.019 – 0.145	0.141	-0.253	-0.888 – 0.382	0.435
Stimulus Presentation Order	0.001	-0.000 – 0.002	0.188	0.001	-0.000 – 0.002	0.194	0.001	-0.000 – 0.002	0.212	0.001	-0.000 – 0.002	0.208	0.001	-0.000 – 0.002	0.180
Task Engagement	-0.189	-0.242 – -0.135	**<0.001**	-0.185	-0.238 – -0.131	**<0.001**	-0.175	-0.229 – -0.121	**<0.001**	-0.166	-0.221 – -0.111	**<0.001**	-0.165	-0.222 – -0.109	**<0.001**
Reminder Group	0.033	-0.057 – 0.123	0.469	0.180	-0.103 – 0.464	0.212	0.365	0.013 – 0.716	**0.042**	0.390	0.037 – 0.742	**0.030**	0.243	-0.174 – 0.659	0.253
Aid Group	-0.633	-0.725 – -0.541	**<0.001**	-0.730	-1.017 – -0.444	**<0.001**	-0.550	-0.906 – -0.193	**0.003**	-0.521	-0.878 – -0.164	**0.004**	-0.607	-1.024 – -0.189	**0.004**
BI x Stimulus Presentation Order	0.000	-0.002 – 0.003	0.707	0.000	-0.002 – 0.003	0.697	0.000	-0.002 – 0.003	0.659	0.000	-0.002 – 0.003	0.665	0.000	-0.002 – 0.003	0.741
BI x Task Engagement	0.290	0.198 – 0.381	**<0.001**	0.287	0.196 – 0.379	**<0.001**	0.263	0.172 – 0.355	**<0.001**	0.241	0.147 – 0.334	**<0.001**	0.242	0.146 – 0.337	**<0.001**
BI x Reminder Group	0.069	-0.082 – 0.221	0.370	0.125	-0.357 – 0.607	0.612	-0.055	-0.649 – 0.539	0.856	-0.122	-0.718 – 0.474	0.688	0.020	-0.678 – 0.717	0.956
BI x Aid Group	1.646	1.489 – 1.803	**<0.001**	1.887	1.399 – 2.376	**<0.001**	1.500	0.894 – 2.105	**<0.001**	1.456	0.851 – 2.062	**<0.001**	1.498	0.798 – 2.197	**<0.001**
Social Media Experience				-0.001	-0.003 – 0.002	0.645	-0.001	-0.003 – 0.002	0.598	-0.000	-0.003 – 0.002	0.684	-0.001	-0.003 – 0.002	0.666
BI x Social Media Experience				0.002	-0.002 – 0.006	0.449	0.002	-0.002 – 0.006	0.368	0.001	-0.002 – 0.005	0.468	0.002	-0.003 – 0.006	0.464
Reminder Group x Social Media Experience				-0.002	-0.005 – 0.001	0.271	-0.002	-0.005 – 0.001	0.234	-0.002	-0.005 – 0.001	0.221	-0.002	-0.006 – 0.001	0.188
Aid Group x Social Media Experience				0.001	-0.002 – 0.005	0.477	0.001	-0.003 – 0.004	0.595	0.001	-0.003 – 0.004	0.696	0.001	-0.003 – 0.004	0.711
BI x Reminder Group x Social Media Experience				-0.001	-0.006 – 0.005	0.823	-0.001	-0.006 – 0.005	0.860	-0.000	-0.006 – 0.005	0.910	-0.000	-0.006 – 0.006	0.978
BI x Aid Group x Social Media Experience				-0.003	-0.009 – 0.003	0.298	-0.002	-0.008 – 0.004	0.433	-0.002	-0.008 – 0.004	0.518	-0.002	-0.008 – 0.004	0.496
Analytical Reasoning (CRT)							0.007	-0.027 – 0.042	0.674	0.013	-0.022 – 0.048	0.462	-0.013	-0.062 – 0.036	0.600
BI x CRT							0.013	-0.046 – 0.071	0.675	0.004	-0.055 – 0.063	0.889	0.007	-0.075 – 0.090	0.859
Reminder Group x CRT							-0.043	-0.092 – 0.006	0.083	-0.051	-0.101 – -0.001	**0.046**	0.000	-0.070 – 0.070	0.995
Aid Group x CRT							-0.041	-0.089 – 0.008	0.098	-0.047	-0.096 – 0.002	0.060	-0.026	-0.095 – 0.043	0.462
BI x Reminder Group x CRT							0.042	-0.040 – 0.125	0.314	0.058	-0.026 – 0.143	0.175	0.019	-0.098 – 0.137	0.746
BI x Aid Group x CRT							0.087	0.005 – 0.169	**0.038**	0.093	0.011 – 0.176	**0.027**	0.082	-0.034 – 0.198	0.164
Participant Political View (PV)										-0.022	-0.097 – 0.053	0.566	0.039	-0.062 – 0.141	0.449
BI x PV										-0.040	-0.166 – 0.087	0.541	-0.156	-0.326 – 0.015	0.074
Reminder Group x PV										0.040	-0.067 – 0.147	0.465	0.078	-0.068 – 0.224	0.293
Aid Group x PV										-0.056	-0.164 – 0.052	0.308	0.008	-0.141 – 0.157	0.919
CRT x PV										0.011	-0.006 – 0.028	0.197	0.021	-0.002 – 0.045	0.076
BI x Reminder Group x PV										-0.018	-0.199 – 0.163	0.846	0.007	-0.238 – 0.252	0.954
BI x Aid Group x PV										0.186	0.003 – 0.369	**0.047**	0.167	-0.085 – 0.418	0.194
BI x CRT x PV										-0.003	-0.032 – 0.025	0.817	-0.003	-0.042 – 0.037	0.892
Reminder Group x PV x CRT										-0.009	-0.034 – 0.015	0.451	-0.007	-0.041 – 0.026	0.668
Aid Group x PV x CRT										0.006	-0.019 – 0.030	0.655	0.001	-0.033 – 0.035	0.955
BI x Reminder Group x PV x CRT										0.007	-0.034 – 0.048	0.740	-0.016	-0.072 – 0.040	0.584
BI x Aid Group x PV x CRT										-0.032	-0.074 – 0.009	0.127	-0.046	-0.103 – 0.011	0.115
N	924			924			924			924			924		
Observations	55440			55440			55440			55440			55440		
Marginal R^2 / Conditional R^2	0.171 / 0.288			0.171 / 0.287			0.172 / 0.287			0.173 / 0.287			0.198 / 0.312		
AUC	0.756			0.756			0.756			0.756			0.764		

Table 1b



	Model 5		
Predictors	*Estimate*	*CI*	*p*
Political Difference between Stimuli and Participant (PD)	0.203	0.046 – 0.360	**0.011**
BI x PD	-0.206	-0.475 – 0.062	0.132
Reminder Group x PD	0.180	-0.052 – 0.412	0.128
Aid Group x PD	0.096	-0.132 – 0.324	0.409
CRT x PD	0.026	-0.011 – 0.063	0.168
PV x PD	-0.066	-0.134 – 0.001	**0.053**
BI x Reminder Group x PD	-0.179	-0.577 – 0.219	0.379
BI x Aid Group x PD	-0.024	-0.416 – 0.368	0.903
BI x CRT x PD	0.000	-0.063 – 0.064	0.989
Reminder Group x CRT x PD	-0.058	-0.112 – -0.004	**0.035**
Aid Group x CRT x PD	-0.028	-0.081 – 0.025	0.305
BI x PV x PD	0.124	0.009 – 0.240	**0.035**
Reminder Group x PV x PD	-0.030	-0.129 – 0.068	0.548
Aid Group x PV x PD	-0.064	-0.166 – 0.037	0.213
CRT x PV x PD	-0.007	-0.023 – 0.009	0.369
BI x Reminder Group x CRT x PD	0.045	-0.047 – 0.137	0.341
BI x Aid Group x CRT x PD	0.019	-0.072 – 0.111	0.680
BI x Reminder Group x PV x PD	-0.038	-0.207 – 0.132	0.662
BI x Aid Group x PV x PD	0.017	-0.158 – 0.192	0.849
BI x CRT x PV x PD	-0.004	-0.031 – 0.024	0.792
Reminder Group x CRT x PV x PD	-0.005	-0.027 – 0.018	0.687
Aid Group x CRT x PV x PD	0.005	-0.018 – 0.028	0.654
BI x Reminder Group x CRT x PV x PD	0.026	-0.013 – 0.065	0.191
BI x Aid Group x CRT x PV x PD	0.013	-0.027 – 0.053	0.523
N		924	
Observations		55440	
Marginal R^2 / Conditional R^2		0.198 / 0.312	
AUC		0.764	

Note. Table 1a and 1b shows results from planned analyses predicting whether participants respond 'bot' = 1, vs. 'human' = 0, as a function of test conditions (control, aid, reminder), task characteristics (bot indicator [BI], stimulus presentation order [SPO]), individual factors (task engagement [TE], social media experience [SEM], cognitive reflection test [CRT], political view [PV]), and political difference (between individual and stimulus; [PD]). The coefficients are estimated with a general linear mixed-effects probit model. The dependent measure was each participant's judgment that a stimulus was a 'bot' or a 'human' persona. BI is the algorithm-derived probability of the stimulus being a bot. The analysis is based on N=928 participants. The intercept reflects the criterion for the average participant in the control condition (with all other regressors at their mean), with higher values indicating a greater tendency to respond 'bot.' Positive interactions with the bot indicator score increased sensitivity (d') for a given subgroup of participants.

Sample Demographics

Nine hundred and thirty-eight participants completed the study. Six participants were excluded for failing to follow Prolific's verification procedures. One was excluded for completing the task too quickly (less than two minutes); three were excluded for taking too long (more than three hours).

The analyzed sample included 928 participants, 373 males, 536 females, 16 non-binary, 3 who preferred not to say; age ranged from 18 to 78 years old (median = 33; mean = 36). Seven hundred and fourteen reported being White, 68 Hispanic or Latino, 49 Black or African American, 5 Native American, 71 Asian or Pacific Islander, and 21 Other. Five reported less than high school, 368 a high school degree or equivalent, 396 a bachelor's degree, 128 a master's degree, and 31 a doctorate. Four hundred and nine reported being fully employed, 123 employed part-time, 89 unemployed but looking, 51 unemployed not looking, 124 students, 35 retired, 76 were self-employed, and 21 unable to work. Three hundred and twenty-eight reported being married, 11 widowed, 70 divorced, 11 separated, and 508 never married. Annual incomes were roughly normally distributed, over 8 categories ranging from "less than 10K" to "over 150K," with the median between 50K

Criterion, Bot Indicator, and Controls

The estimated SDT parameters assume that the signal and noise distributions have equal variance and are Gaussian (Lynn & Barrett, 2014). A log-linear correction added 0.5 to the number of hits and false alarms and 1 to the number of signals (bot personas) or noise (human personas) to correct for participants who identified all stimuli correctly or incorrectly, thereby producing hit (H) or false alarm (FA) rates of 0 or 1 (Hautus, 1995). Thus, d' and c were calculated using:

$$H = (hits + 0.5)/(signals + 1)$$

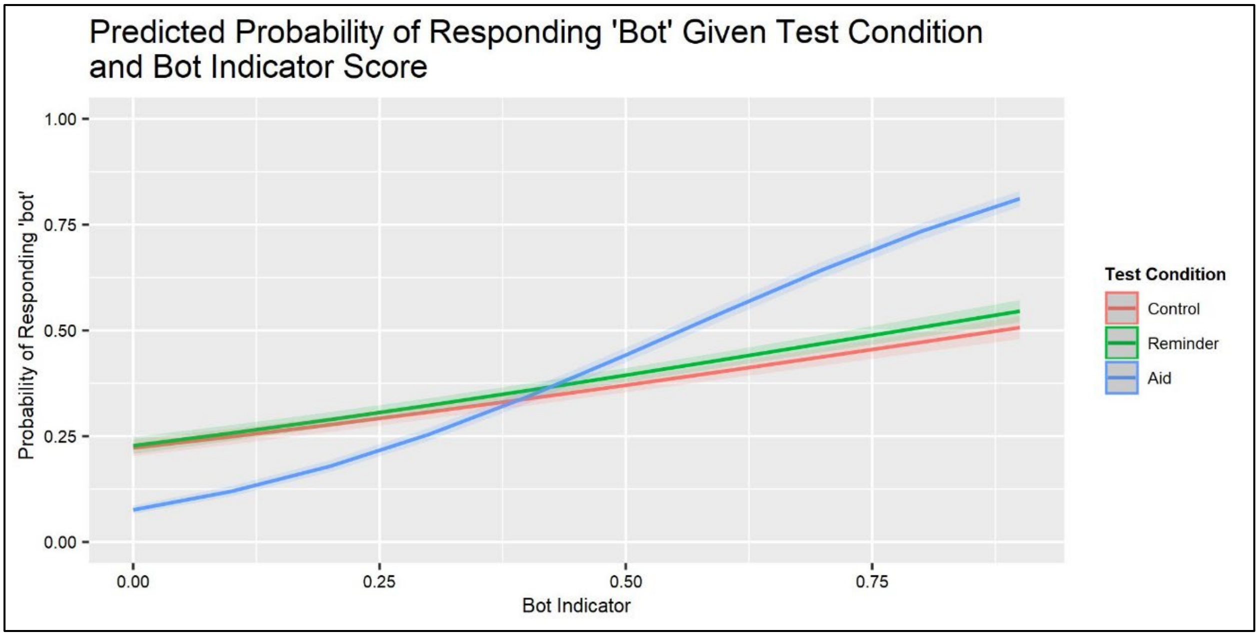
$$FA = (false\,alarms + 0.5)/(noise + 1)$$

$$d'\ (sensitivity) = z(H) - z(FA)$$

$$c\ (criterion) = -0.5[z(H) + z(FA)]$$

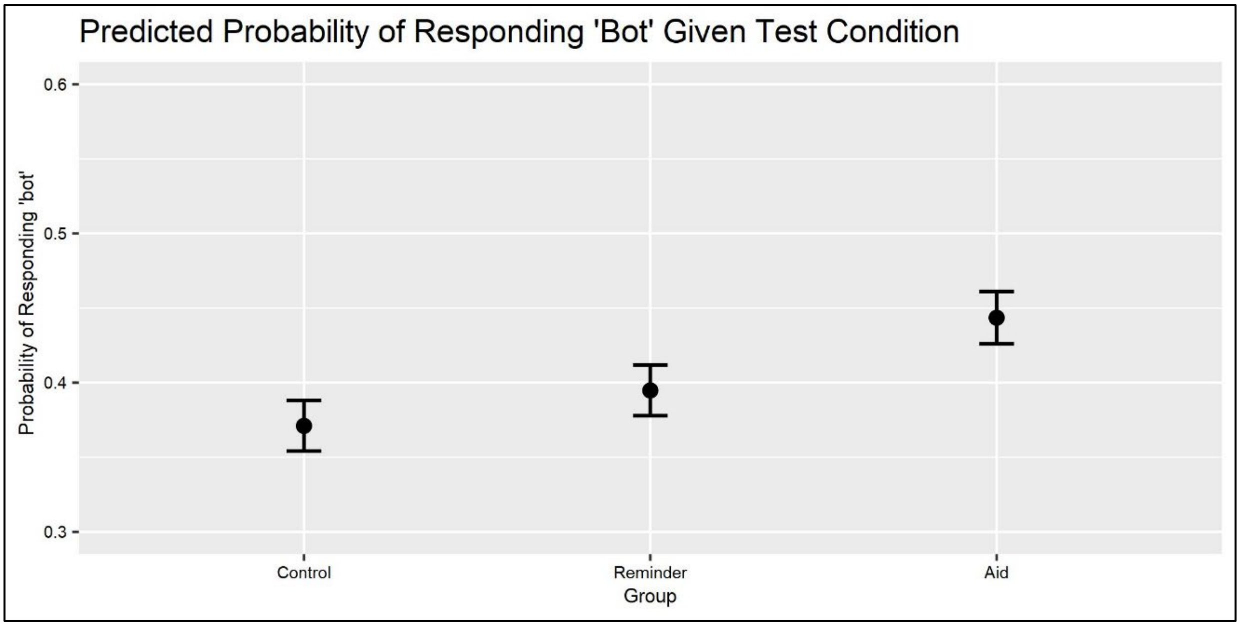
Average Sensitivity. Figure 3 shows functions predicting the probability of responding “bot” based on BI, with the covariates in Model 5. Participants in all three conditions demonstrated sensitivity to the bot indicator score, as evidenced by the positive trend of the curve, and the significant BI x Test Condition coefficient in Model 5. However, there were significant differences across conditions ($p < 0.001$), with participants in the Aid condition showing greater sensitivity than those in the Control or Reminder Group – which were indistinguishable from one another and very similar to the control group in Study 1.

Figure 3



Average Criterion. Figure 4 shows the distributions of criteria for participants in the three conditions. Each group had criteria reflecting a tendency to respond “human” (and hesitancy to respond 'bot'). Participants in the aid condition had significantly higher criteria, reflecting a greater willingness to call a persona a bot (p = 0.004).

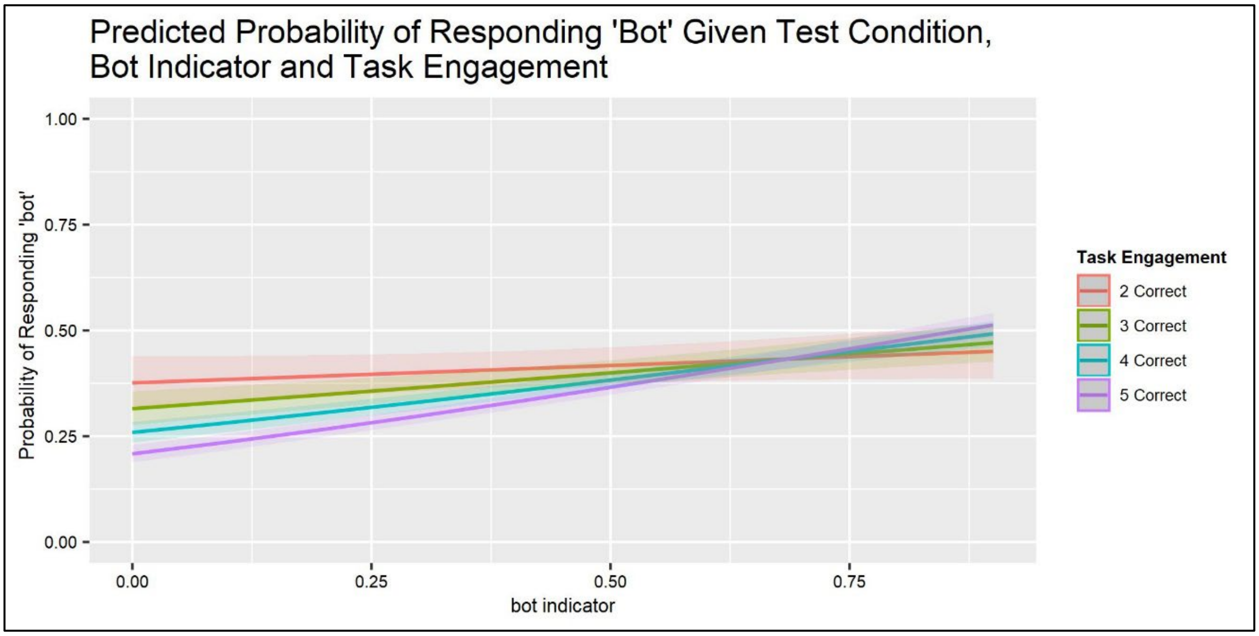
Figure 4



Note. A probability of 0.50 indicates no tendency to respond 'bot' or 'human.' Lower probabilities indicate a tendency to treat a persona as a human and require stronger evidence to treat it as a bot.

Controls. Stimulus presentation order was unrelated to participants' criterion or sensitivity (Model 1, Rows 3, and 7, respectively), contrary to H12a and consistent with H12c. Higher task engagement was associated with stricter criteria (p < 0.001) and greater sensitivity (both p < 0.001; Model 1, Rows 4 and 8, respectively), consistent with H12b, but not H12c. Figure 5 shows these relationships.

Figure 5



Note. Task engagement was measured as the total number of correct attention checks (one prior, three during the experimental procedure, and one post).

Social Media Experience

Model 2 adds Social Media Experience (SME) to the model. It was unrelated to participants' criteria or sensitivity (first and second additional rows, respectively). Social media experience did not figure in any significant interaction or change the relationships in Model 1. Thus, we did not replicate the result from Study 1 or confirm H11.

Analytical Reasoning

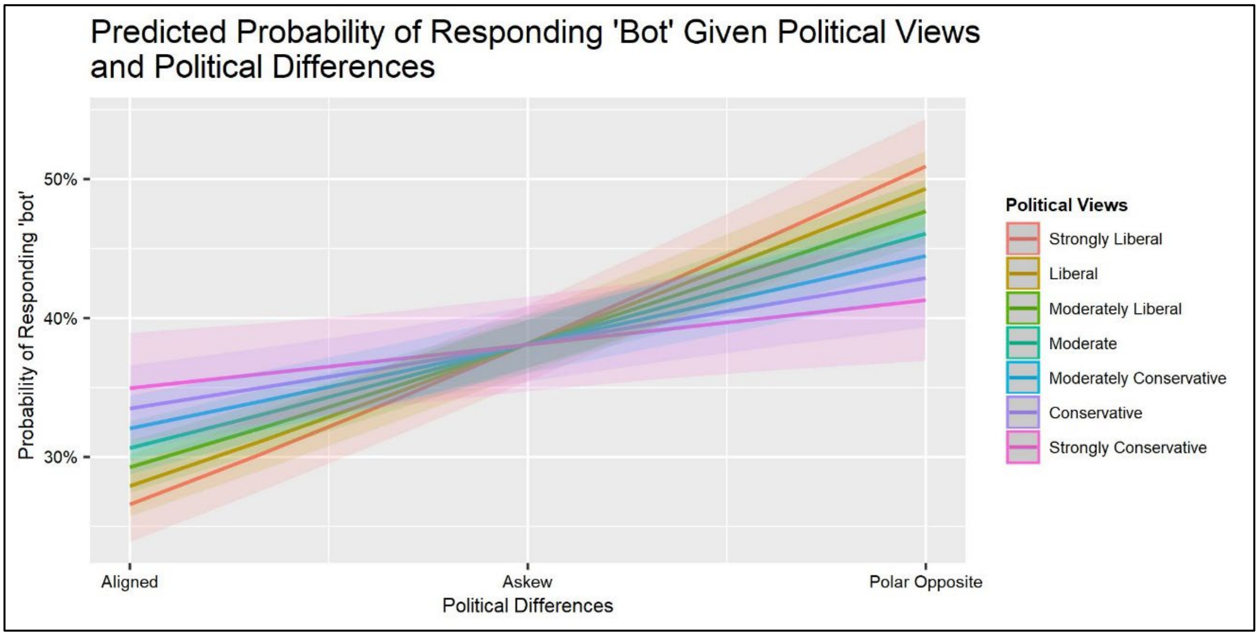
Model 3 adds Cognitive Reflection Test (CRT) scores to the model. They were unrelated to criterion or sensitivity (first and second additional rows, respectively), failing to confirm H9. These scores did not change the relationships in Models 1 or 2. They figured in one weakly significant and hard-to-interpret three-way interaction.

Political Values and Political Differences

Model 4 adds participants' political views (PV). As in Study 1, they were unrelated to criterion or sensitivity (first and second additional rows, respectively). They made little change to the relationships in Models 1-3. They figured in one weakly significant and hard-to-interpret three-way interaction. These findings did not vary by test condition,

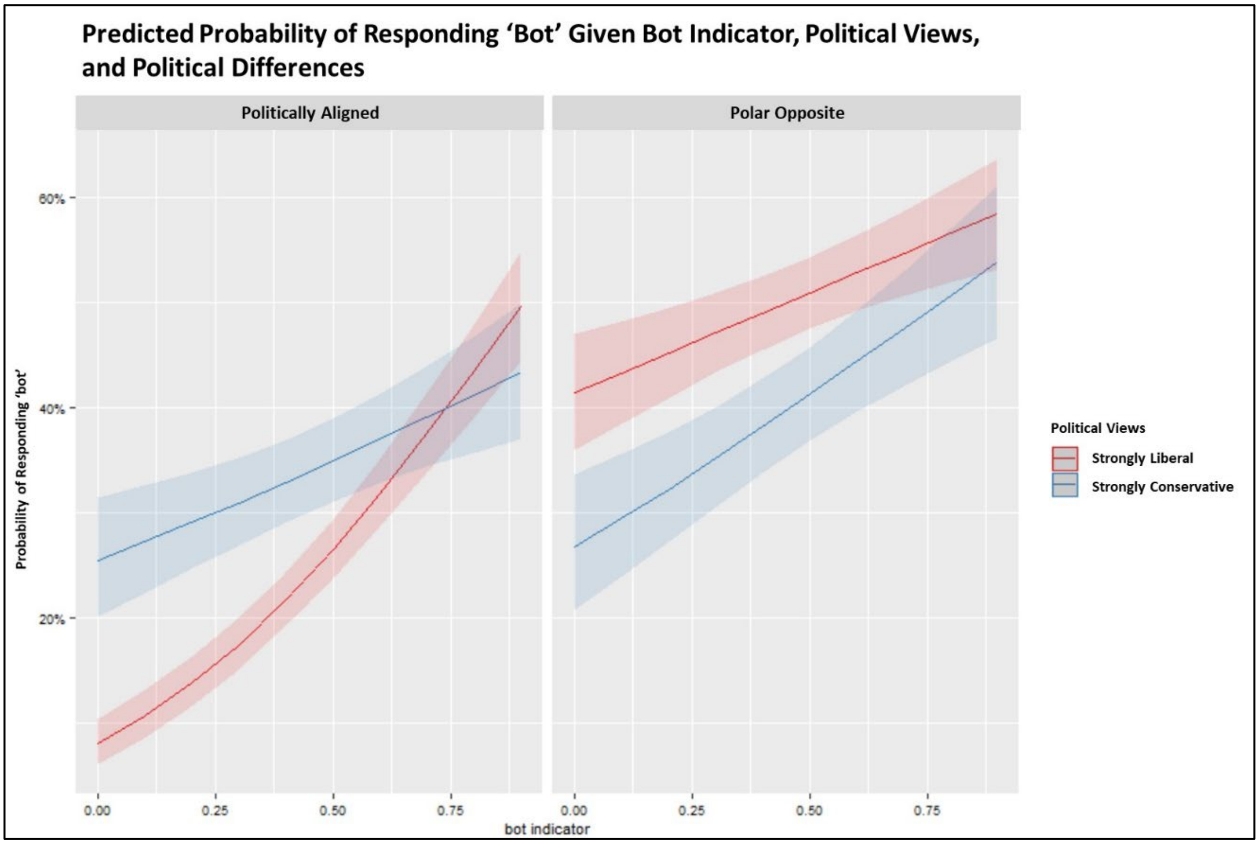
Model 5 adds the political difference (PD) between participants' views and those of the stimulus persona. Consistent with myside bias (and H7a), there was a significant PD main effect, reflecting participants' greater tendency to treat a persona as a human when they agreed with its political view. As in Study 1 (and predicted in H7b), this shift differed for liberal and conservative participants, with liberals' thresholds for responding 'bot' when viewing conservatives decreasing to a greater extent than conservatives' thresholds when considering liberals (reflected in the PD x PV interaction). Figure 6 depicts this relationship. As expected, these findings did not vary by test condition.

Figure 6



Political difference was unrelated to sensitivity alone (as seen in the non-significant BI x PD interaction). Thus , participants were no more (or less) discerning when evaluating personas with similar or different political views. However, as seen in Figure 7, there was significant three-way interaction (BI x PV x PD) (p = 0.035), revealing an asymmetric sensitivity pattern between liberals and conservatives. Conservatives judging liberals (right plot) were modestly more sensitive to the bot indicator score than when they were considering other conservatives (left side). Conversely, liberals judging liberals (left side) were more sensitive to the bot indicator score than liberals considering conservatives (right side). As expected, these findings did not vary by test condition.

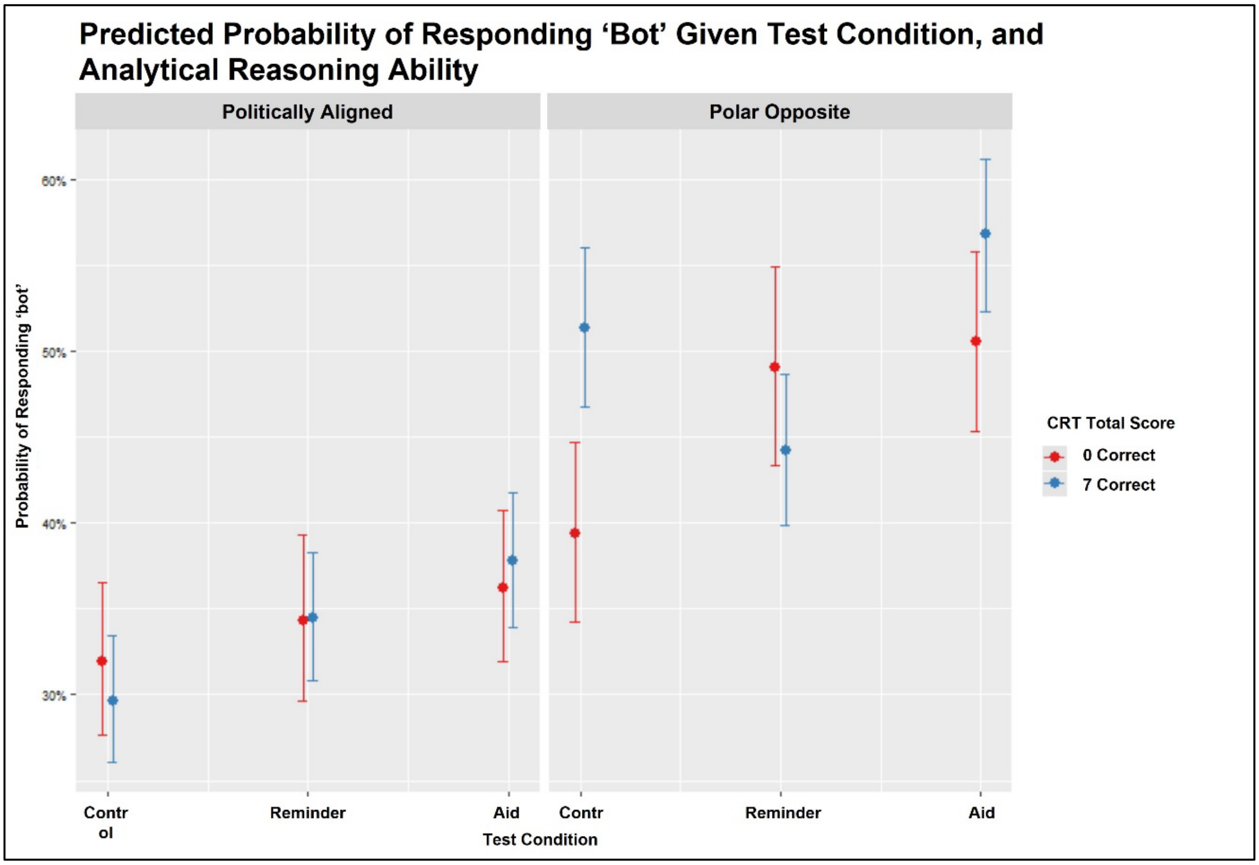
Figure 7



Note. When PD = 0, a persona's political tone matches the participant's self-reported political views. When PD = 2, the views of the persona and participant are opposites.

CRT scores figured in a significant three-way interaction (Reminder Condition x CRT x PD) (p = 0.035). As seen in Figure 8, in the aid and control conditions, but not the reminder, participants with higher analytical reasoning had more lenient criteria when there were political differences. As expected, these findings did not vary by test condition.

Figure 8



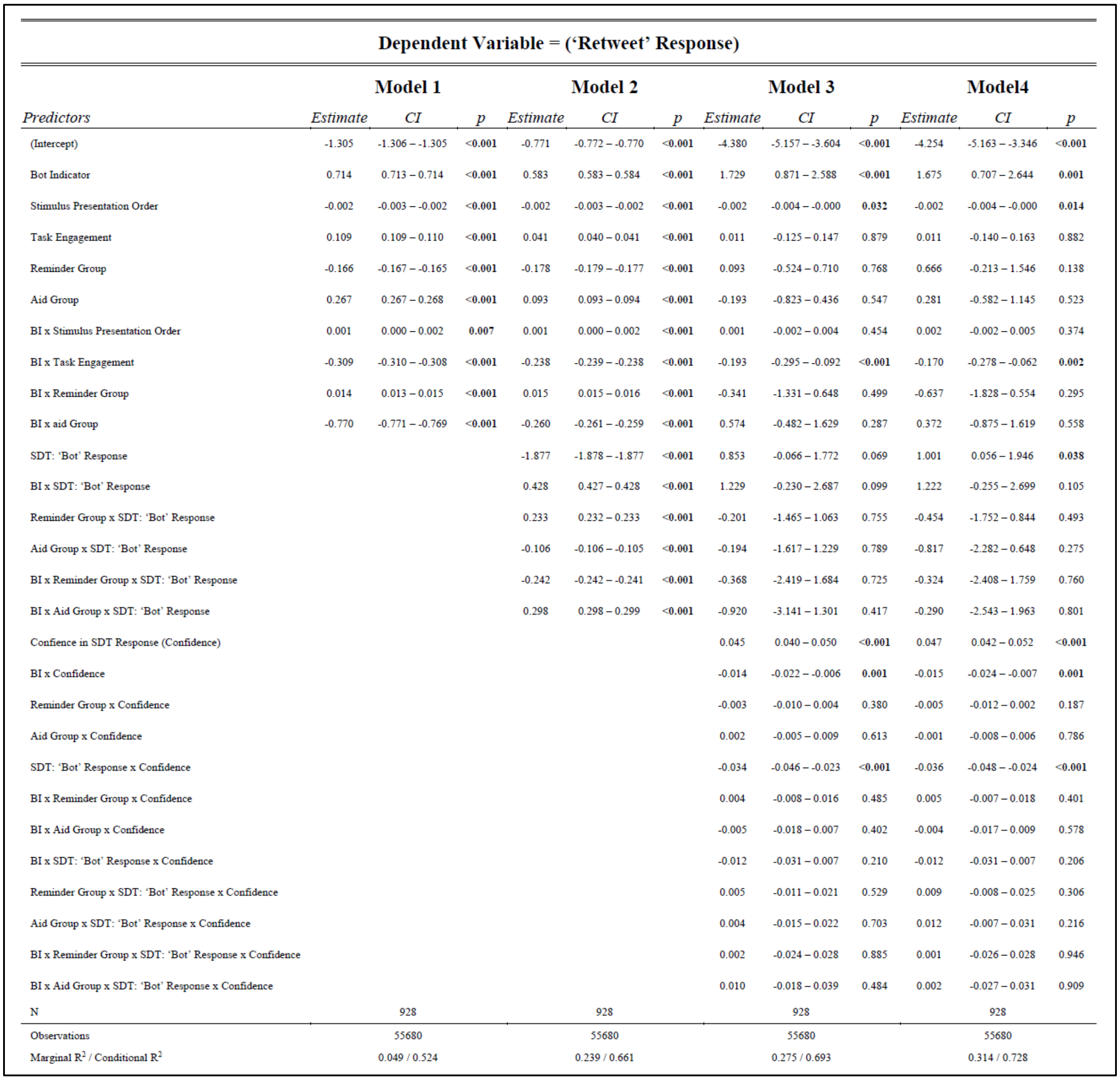
Note. In 'politically aligned' trials, PD= 0. In 'polar opposite' trials, PD=2. The CRT Total Scores are on a scale from 0 correct to 7 correct. The analysis (in Model 5) included all responses; the figure only considers the extremes (0,7) for display purposes.

Behavioral Responses

Willingness to retweet. Tables 2a and 2b present analyses predicting participants' willingness to retweet content from a persona if they agreed with its content. We used the same general linear mixed-effects modeling process as in Tables 1a and 1b, only now to determine the probability of the binary outcome of 'retweet' versus no retweet. As seen in Figures 10 and 11 (whose accompanying analyses are explained below), participants rarely said that they would retweet content.

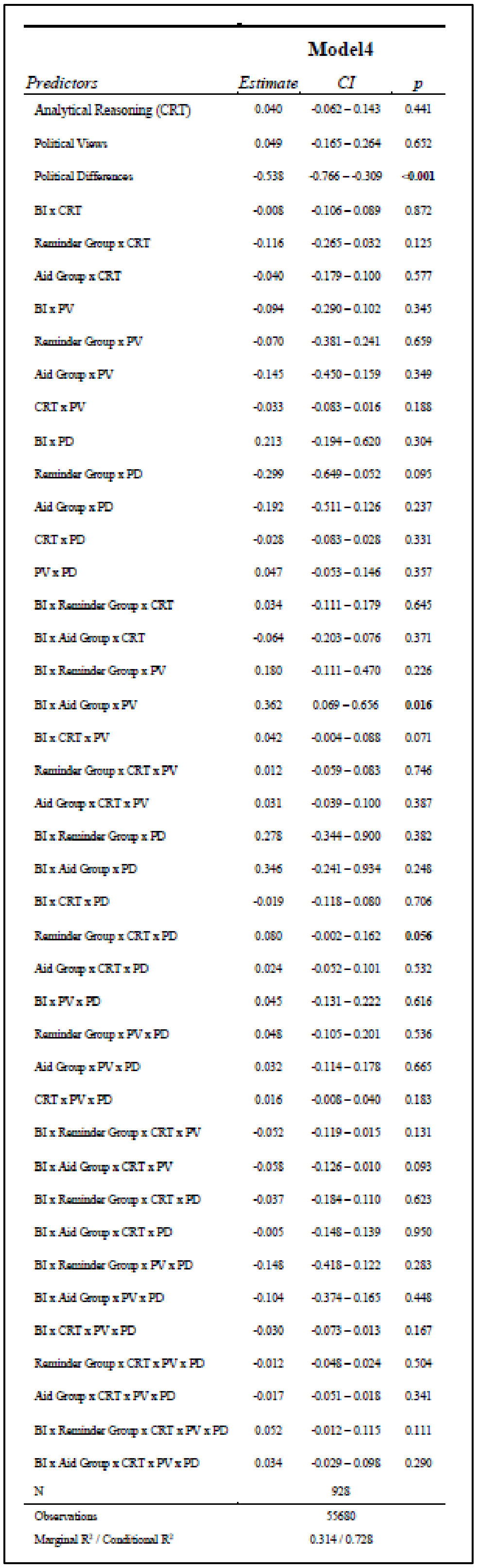
Table 2a

General Linear Mixed Effects Probit Regression Models, Predicting the Probability of Retweeting Content from a Persona



Dependent Variable = ('Retweet' Response)												
	Model 1			Model 2			Model 3			Model4		
Predictors	*Estimate*	*CI*	*p*	*Estimate*	*CI*	*p*	*Estimate*	*CI*	*p*	*Estimate*	*CI*	*p*
(Intercept)	-1.305	-1.306 – -1.305	**<0.001**	-0.771	-0.772 – -0.770	**<0.001**	-4.380	-5.157 – -3.604	**<0.001**	-4.254	-5.163 – -3.346	**<0.001**
Bot Indicator	0.714	0.713 – 0.714	**<0.001**	0.583	0.583 – 0.584	**<0.001**	1.729	0.871 – 2.588	**<0.001**	1.675	0.707 – 2.644	**0.001**
Stimulus Presentation Order	-0.002	-0.003 – -0.002	**<0.001**	-0.002	-0.003 – -0.002	**<0.001**	-0.002	-0.004 – -0.000	**0.032**	-0.002	-0.004 – -0.000	**0.014**
Task Engagement	0.109	0.109 – 0.110	**<0.001**	0.041	0.040 – 0.041	**<0.001**	0.011	-0.125 – 0.147	0.879	0.011	-0.140 – 0.163	0.882
Reminder Group	-0.166	-0.167 – -0.165	**<0.001**	-0.178	-0.179 – -0.177	**<0.001**	0.093	-0.524 – 0.710	0.768	0.666	-0.213 – 1.546	0.138
Aid Group	0.267	0.267 – 0.268	**<0.001**	0.093	0.093 – 0.094	**<0.001**	-0.193	-0.823 – 0.436	0.547	0.281	-0.582 – 1.145	0.523
BI x Stimulus Presentation Order	0.001	0.000 – 0.002	**0.007**	0.001	0.000 – 0.002	**<0.001**	0.001	-0.002 – 0.004	0.454	0.002	-0.002 – 0.005	0.374
BI x Task Engagement	-0.309	-0.310 – -0.308	**<0.001**	-0.238	-0.239 – -0.238	**<0.001**	-0.193	-0.295 – -0.092	**<0.001**	-0.170	-0.278 – -0.062	**0.002**
BI x Reminder Group	0.014	0.013 – 0.015	**<0.001**	0.015	0.015 – 0.016	**<0.001**	-0.341	-1.331 – 0.648	0.499	-0.637	-1.828 – 0.554	0.295
BI x aid Group	-0.770	-0.771 – -0.769	**<0.001**	-0.260	-0.261 – -0.259	**<0.001**	0.574	-0.482 – 1.629	0.287	0.372	-0.875 – 1.619	0.558
SDT: 'Bot' Response				-1.877	-1.878 – -1.877	**<0.001**	0.853	-0.066 – 1.772	0.069	1.001	0.056 – 1.946	**0.038**
BI x SDT: 'Bot' Response				0.428	0.427 – 0.428	**<0.001**	1.229	-0.230 – 2.687	0.099	1.222	-0.255 – 2.699	0.105
Reminder Group x SDT: 'Bot' Response				0.233	0.232 – 0.233	**<0.001**	-0.201	-1.465 – 1.063	0.755	-0.454	-1.752 – 0.844	0.493
Aid Group x SDT: 'Bot' Response				-0.106	-0.106 – -0.105	**<0.001**	-0.194	-1.617 – 1.229	0.789	-0.817	-2.282 – 0.648	0.275
BI x Reminder Group x SDT: 'Bot' Response				-0.242	-0.242 – -0.241	**<0.001**	-0.368	-2.419 – 1.684	0.725	-0.324	-2.408 – 1.759	0.760
BI x Aid Group x SDT: 'Bot' Response				0.298	0.298 – 0.299	**<0.001**	-0.920	-3.141 – 1.301	0.417	-0.290	-2.543 – 1.963	0.801
Confience in SDT Response (Confidence)							0.045	0.040 – 0.050	**<0.001**	0.047	0.042 – 0.052	**<0.001**
BI x Confidence							-0.014	-0.022 – -0.006	**0.001**	-0.015	-0.024 – -0.007	**0.001**
Reminder Group x Confidence							-0.003	-0.010 – 0.004	0.380	-0.005	-0.012 – 0.002	0.187
Aid Group x Confidence							0.002	-0.005 – 0.009	0.613	-0.001	-0.008 – 0.006	0.786
SDT: 'Bot' Response x Confidence							-0.034	-0.046 – -0.023	**<0.001**	-0.036	-0.048 – -0.024	**<0.001**
BI x Reminder Group x Confidence							0.004	-0.008 – 0.016	0.485	0.005	-0.007 – 0.018	0.401
BI x Aid Group x Confidence							-0.005	-0.018 – 0.007	0.402	-0.004	-0.017 – 0.009	0.578
BI x SDT: 'Bot' Response x Confidence							-0.012	-0.031 – 0.007	0.210	-0.012	-0.031 – 0.007	0.206
Reminder Group x SDT: 'Bot' Response x Confidence							0.005	-0.011 – 0.021	0.529	0.009	-0.008 – 0.025	0.306
Aid Group x SDT: 'Bot' Response x Confidence							0.004	-0.015 – 0.022	0.703	0.012	-0.007 – 0.031	0.216
BI x Reminder Group x SDT: 'Bot' Response x Confidence							0.002	-0.024 – 0.028	0.885	0.001	-0.026 – 0.028	0.946
BI x Aid Group x SDT: 'Bot' Response x Confidence							0.010	-0.018 – 0.039	0.484	0.002	-0.027 – 0.031	0.909
N		928			928			928			928	
Observations		55680			55680			55680			55680	
Marginal R^2 / Conditional R^2		0.049 / 0.524			0.239 / 0.661			0.275 / 0.693			0.314 / 0.728	

Table 2b



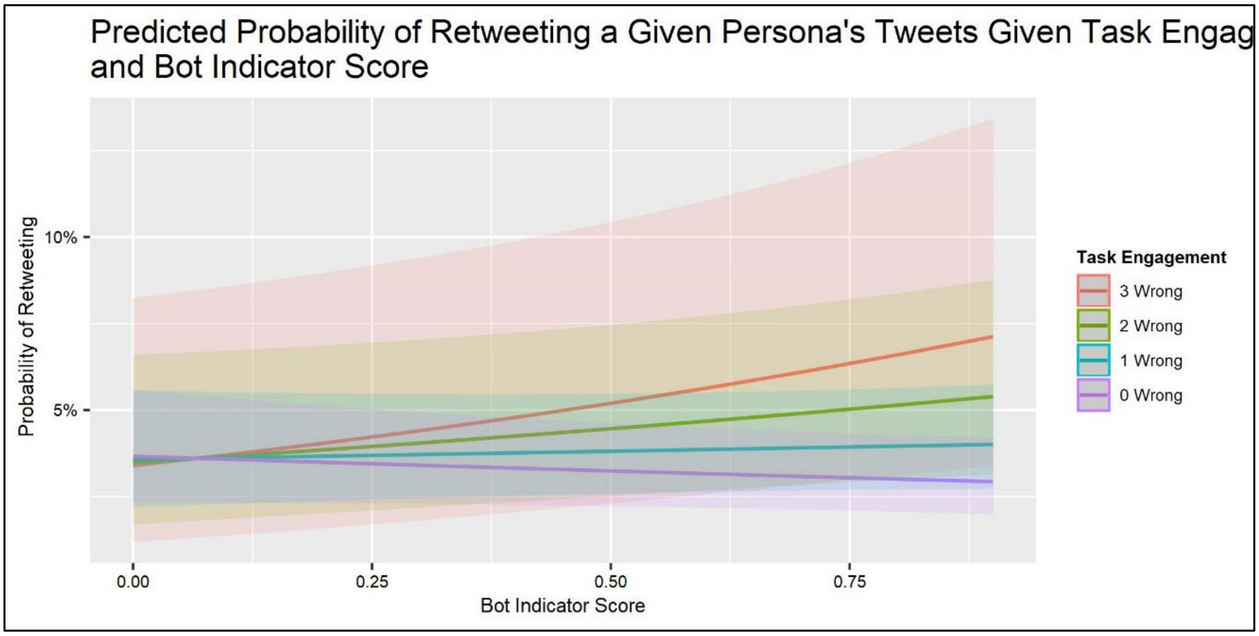
	Model4		
Predictors	*Estimate*	*CI*	*p*
Analytical Reasoning (CRT)	0.040	-0.062 – 0.143	0.441
Political Views	0.049	-0.165 – 0.264	0.652
Political Differences	-0.538	-0.766 – -0.309	**<0.001**
BI x CRT	-0.008	-0.106 – 0.089	0.872
Reminder Group x CRT	-0.116	-0.265 – 0.032	0.125
Aid Group x CRT	-0.040	-0.179 – 0.100	0.577
BI x PV	-0.094	-0.290 – 0.102	0.345
Reminder Group x PV	-0.070	-0.381 – 0.241	0.659
Aid Group x PV	-0.145	-0.450 – 0.159	0.349
CRT x PV	-0.033	-0.083 – 0.016	0.188
BI x PD	0.213	-0.194 – 0.620	0.304
Reminder Group x PD	-0.299	-0.649 – 0.052	0.095
Aid Group x PD	-0.192	-0.511 – 0.126	0.237
CRT x PD	-0.028	-0.083 – 0.028	0.331
PV x PD	0.047	-0.053 – 0.146	0.357
BI x Reminder Group x CRT	0.034	-0.111 – 0.179	0.645
BI x Aid Group x CRT	-0.064	-0.203 – 0.076	0.371
BI x Reminder Group x PV	0.180	-0.111 – 0.470	0.226
BI x Aid Group x PV	0.362	0.069 – 0.656	**0.016**
BI x CRT x PV	0.042	-0.004 – 0.088	0.071
Reminder Group x CRT x PV	0.012	-0.059 – 0.083	0.746
Aid Group x CRT x PV	0.031	-0.039 – 0.100	0.387
BI x Reminder Group x PD	0.278	-0.344 – 0.900	0.382
BI x Aid Group x PD	0.346	-0.241 – 0.934	0.248
BI x CRT x PD	-0.019	-0.118 – 0.080	0.706
Reminder Group x CRT x PD	0.080	-0.002 – 0.162	**0.056**
Aid Group x CRT x PD	0.024	-0.052 – 0.101	0.532
BI x PV x PD	0.045	-0.131 – 0.222	0.616
Reminder Group x PV x PD	0.048	-0.105 – 0.201	0.536
Aid Group x PV x PD	0.032	-0.114 – 0.178	0.665
CRT x PV x PD	0.016	-0.008 – 0.040	0.183
BI x Reminder Group x CRT x PV	-0.052	-0.119 – 0.015	0.131
BI x Aid Group x CRT x PV	-0.058	-0.126 – 0.010	0.093
BI x Reminder Group x CRT x PD	-0.037	-0.184 – 0.110	0.623
BI x Aid Group x CRT x PD	-0.005	-0.148 – 0.139	0.950
BI x Reminder Group x PV x PD	-0.148	-0.418 – 0.122	0.283
BI x Aid Group x PV x PD	-0.104	-0.374 – 0.165	0.448
BI x CRT x PV x PD	-0.030	-0.073 – 0.013	0.167
Reminder Group x CRT x PV x PD	-0.012	-0.048 – 0.024	0.504
Aid Group x CRT x PV x PD	-0.017	-0.051 – 0.018	0.341
BI x Reminder Group x CRT x PV x PD	0.052	-0.012 – 0.115	0.111
BI x Aid Group x CRT x PV x PD	0.034	-0.029 – 0.098	0.290
N	928		
Observations	55680		
Marginal R^2 / Conditional R^2	0.314 / 0.728		

Model 1 shows predictions based on experimental condition, stimulus BI, the two control variables (stimulus presentation order, task engagement), and their interactions. Model 2 adds participants' responses (bot or human) and interactions. Model 3 adds their confidence in that response. Model 4 adds PV, PD, and CRT. As social media experience had no predictive value for sensitivity or criterion, or interaction with any variable, we made the post hoc decision to drop it from subsequent analyses.

Model 1 shows that participants were significantly more likely to retweet messages the higher a persona's bot indicator score, consistent with sensitivity. The model contrasts the aid and reminder groups with the control group. Willingness to retweet was higher in the Aid condition (positive coefficient), suggesting that it increased their confidence in retweeting, and lower in the Reminder group (negative coefficient), indicating that it decreased participants' confidence (by giving a warning with no instruction for acting on it). Significant interactions indicated that BI was less valuable for the Reminder group and more valuable for the Aid group.

Retweeting declined as the experiment progressed (and SPO increased), suggesting fatigue. Figure 9 reveals it was more common for participants with greater task engagement, implying that retweet decisions require effort. Both control variables had interactions with BI, suggesting that participants who were more engaged and less tired could extract more information from the personas.

Figure 9

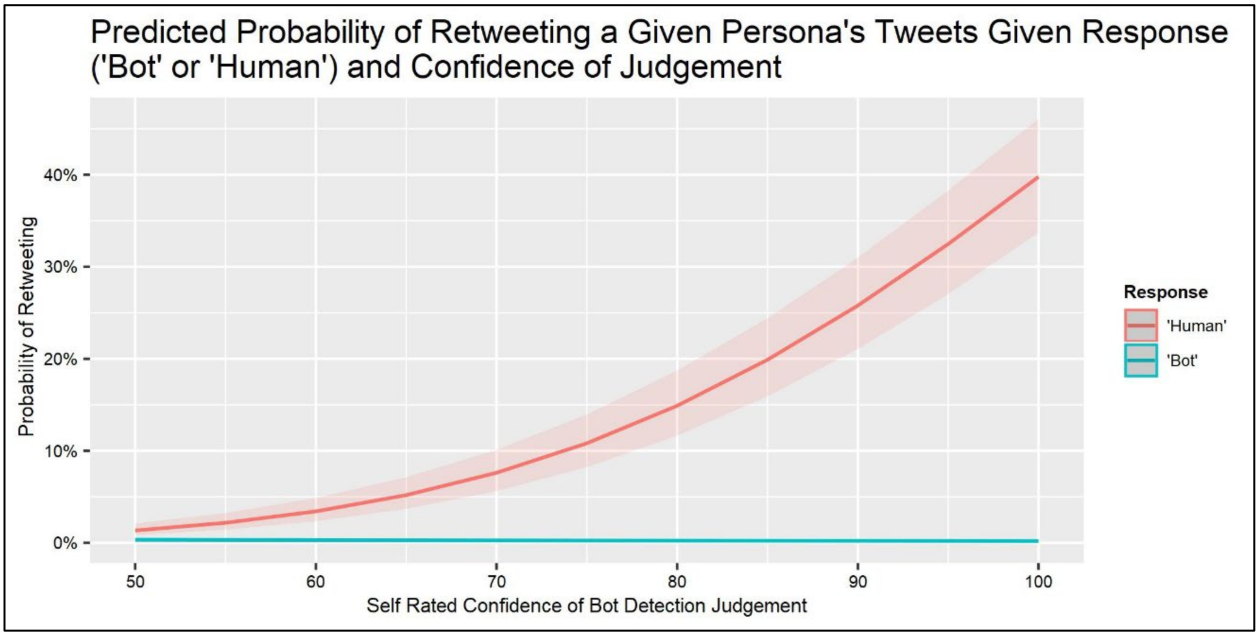


Model 2 added participants' responses (bot or human) to each persona. Not surprisingly, participants were more willing to tweet content from personas they judged to be humans. That tendency increased with BI (BIxSDT interaction).

Overall, participants expressed moderate confidence in their bot/human judgments. Pooling groups and stimuli: M=77.8% (SD=11.5%), on the 50-100% scale. Across the test conditions, average confidence did not vary significantly (control = 77.8%, reminder = 77.1%, aid = 78.5%). Overall, 61.9% of participants' judgments were accurate. As seen in the sensitivity findings, accuracy did vary by test condition (control = 56.3%, reminder = 57.3%, aid = 71.9%). Thus, participants were overconfident in their abilities, replicating a familiar finding, in this novel setting (Canfield et al., 2019; Lichtenstein et al., 1982). However, overconfidence was much less in the aid condition (6.6%) than the other two (control = 21.5%; remainder = 19.8%)

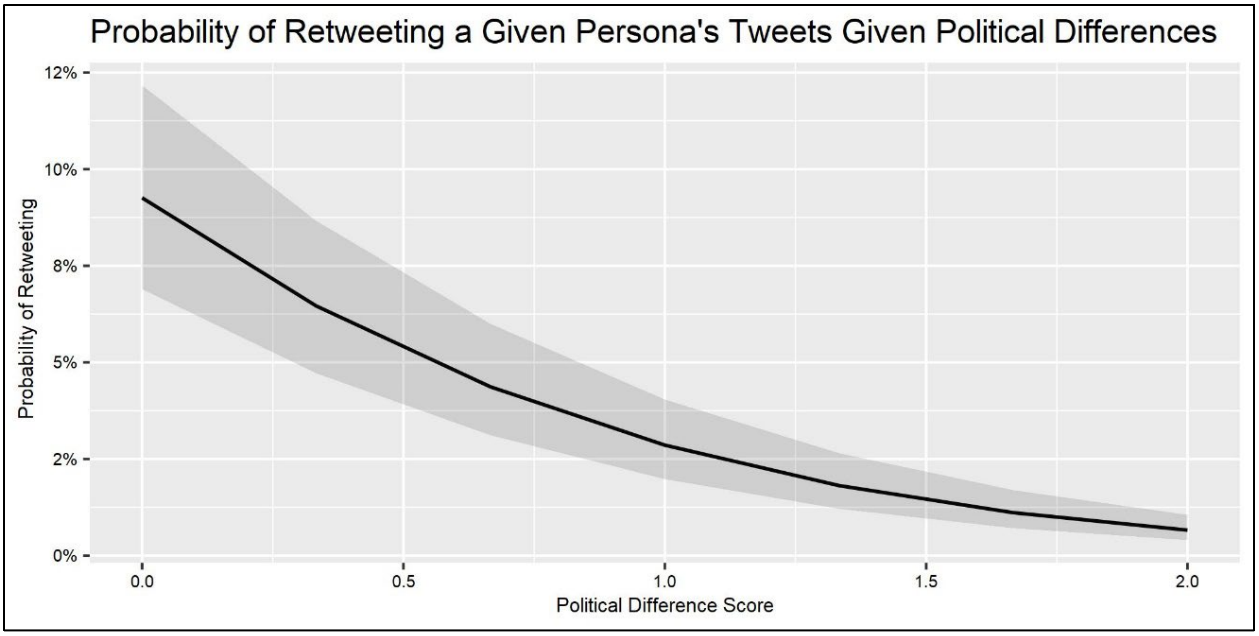
Model 3 added participants' confidence in their response (bot or human) and related interactions. Confidence proved to be a powerful predictor of willingness to retweet. Interactions showed that BI scores amplified the importance of confidence while responding "bot" decreased it. Indeed, once confidence is included in the model, the only other predictors of willingness to retweet are BI and the interaction between BI and Task Engagement. Figure 10 depicts these patterns, pooling the three conditions, which did not differ.

Figure 10



Model 4 adds CRT, PV, PD, and the related interactions. The only significant predictor was that participants were much less likely to retweet content they agreed with if it came from a persona with different political views (see Figure 11). Model 3 patterns did not change with the addition of these predictors. Here, too, test conditions did not matter.

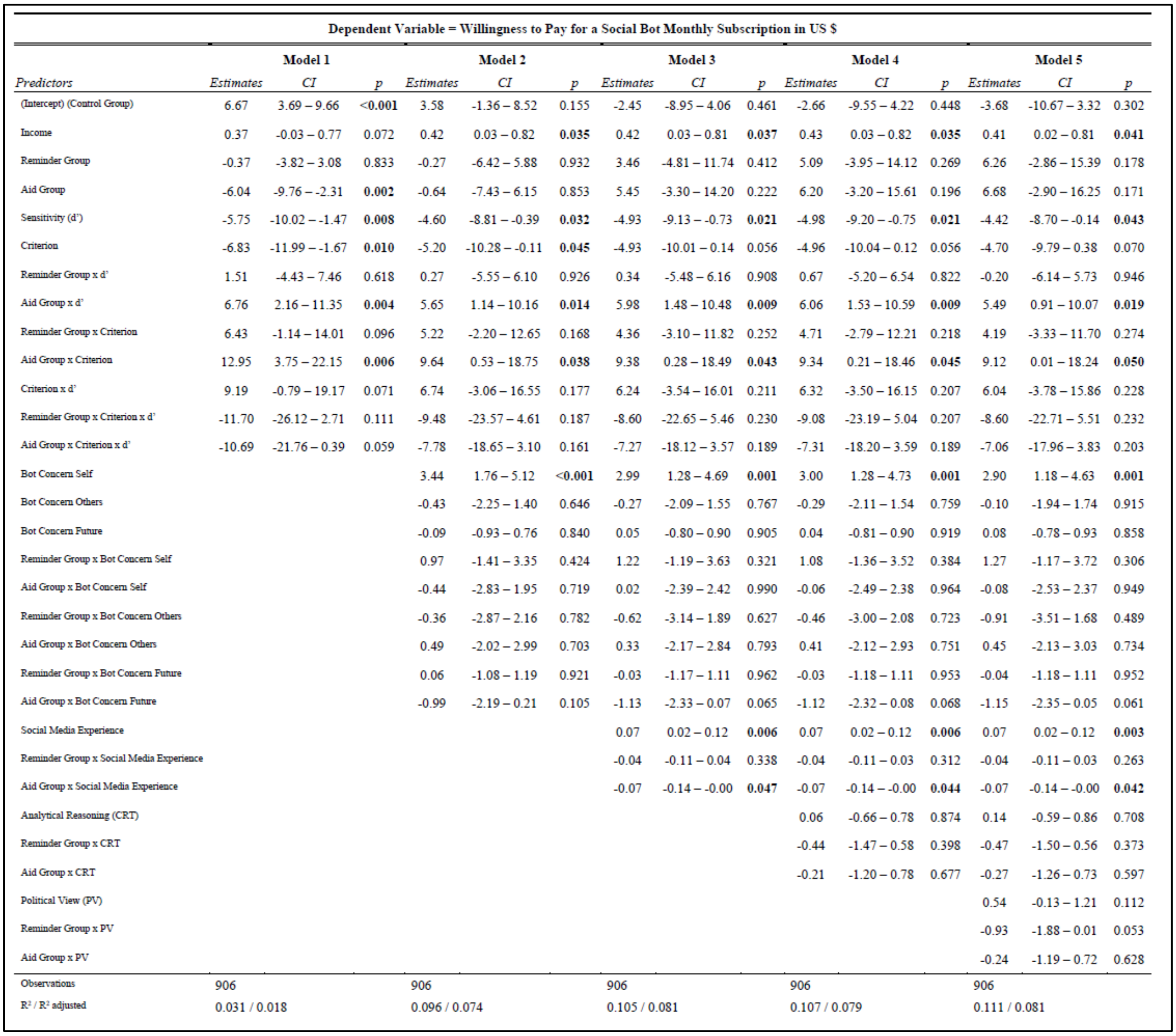
Figure 11



Willingness to Pay – Monetary. We pursued a similar analytical strategy for participants' other behavioral responses, their monthly willingness-to-pay (WTP) for an automated social bot detection advisor subscription. Participants could enter any amount, with no upper bounds, rounded to $1 intervals. Table 3 in SM presents detailed results. All models included income, finding that participants who reported having more were also willing to pay somewhat more.

Table 3

Linear Models Predicting Willingness to Pay for a Monthly Social Bot Detection Service.



	Dependent Variable = Willingness to Pay for a Social Bot Monthly Subscription in US $														
		Model 1			**Model 2**			**Model 3**			**Model 4**			**Model 5**	
Predictors	*Estimates*	*CI*	*p*	*Estimates*	*CI*	*p*	*Estimates*	*CI*	*p*	*Estimates*	*CI*	*p*	*Estimates*	*CI*	*p*
(Intercept) (Control Group)	6.67	3.69 – 9.66	**<0.001**	3.58	-1.36 – 8.52	0.155	-2.45	-8.95 – 4.06	0.461	-2.66	-9.55 – 4.22	0.448	-3.68	-10.67 – 3.32	0.302
Income	0.37	-0.03 – 0.77	0.072	0.42	0.03 – 0.82	**0.035**	0.42	0.03 – 0.81	**0.037**	0.43	0.03 – 0.82	**0.035**	0.41	0.02 – 0.81	**0.041**
Reminder Group	-0.37	-3.82 – 3.08	0.833	-0.27	-6.42 – 5.88	0.932	3.46	-4.81 – 11.74	0.412	5.09	-3.95 – 14.12	0.269	6.26	-2.86 – 15.39	0.178
Aid Group	-6.04	-9.76 – -2.31	**0.002**	-0.64	-7.43 – 6.15	0.853	5.45	-3.30 – 14.20	0.222	6.20	-3.20 – 15.61	0.196	6.68	-2.90 – 16.25	0.171
Sensitivity (d')	-5.75	-10.02 – -1.47	**0.008**	-4.60	-8.81 – -0.39	**0.032**	-4.93	-9.13 – -0.73	**0.021**	-4.98	-9.20 – -0.75	**0.021**	-4.42	-8.70 – -0.14	**0.043**
Criterion	-6.83	-11.99 – -1.67	**0.010**	-5.20	-10.28 – -0.11	**0.045**	-4.93	-10.01 – 0.14	0.056	-4.96	-10.04 – 0.12	0.056	-4.70	-9.79 – 0.38	0.070
Reminder Group x d'	1.51	-4.43 – 7.46	0.618	0.27	-5.55 – 6.10	0.926	0.34	-5.48 – 6.16	0.908	0.67	-5.20 – 6.54	0.822	-0.20	-6.14 – 5.73	0.946
Aid Group x d'	6.76	2.16 – 11.35	**0.004**	5.65	1.14 – 10.16	**0.014**	5.98	1.48 – 10.48	**0.009**	6.06	1.53 – 10.59	**0.009**	5.49	0.91 – 10.07	**0.019**
Reminder Group x Criterion	6.43	-1.14 – 14.01	0.096	5.22	-2.20 – 12.65	0.168	4.36	-3.10 – 11.82	0.252	4.71	-2.79 – 12.21	0.218	4.19	-3.33 – 11.70	0.274
Aid Group x Criterion	12.95	3.75 – 22.15	**0.006**	9.64	0.53 – 18.75	**0.038**	9.38	0.28 – 18.49	**0.043**	9.34	0.21 – 18.46	**0.045**	9.12	0.01 – 18.24	**0.050**
Criterion x d'	9.19	-0.79 – 19.17	0.071	6.74	-3.06 – 16.55	0.177	6.24	-3.54 – 16.01	0.211	6.32	-3.50 – 16.15	0.207	6.04	-3.78 – 15.86	0.228
Reminder Group x Criterion x d'	-11.70	-26.12 – 2.71	0.111	-9.48	-23.57 – 4.61	0.187	-8.60	-22.65 – 5.46	0.230	-9.08	-23.19 – 5.04	0.207	-8.60	-22.71 – 5.51	0.232
Aid Group x Criterion x d'	-10.69	-21.76 – 0.39	0.059	-7.78	-18.65 – 3.10	0.161	-7.27	-18.12 – 3.57	0.189	-7.31	-18.20 – 3.59	0.189	-7.06	-17.96 – 3.83	0.203
Bot Concern Self				3.44	1.76 – 5.12	**<0.001**	2.99	1.28 – 4.69	**0.001**	3.00	1.28 – 4.73	**0.001**	2.90	1.18 – 4.63	**0.001**
Bot Concern Others				-0.43	-2.25 – 1.40	0.646	-0.27	-2.09 – 1.55	0.767	-0.29	-2.11 – 1.54	0.759	-0.10	-1.94 – 1.74	0.915
Bot Concern Future				-0.09	-0.93 – 0.76	0.840	0.05	-0.80 – 0.90	0.905	0.04	-0.81 – 0.90	0.919	0.08	-0.78 – 0.93	0.858
Reminder Group x Bot Concern Self				0.97	-1.41 – 3.35	0.424	1.22	-1.19 – 3.63	0.321	1.08	-1.36 – 3.52	0.384	1.27	-1.17 – 3.72	0.306
Aid Group x Bot Concern Self				-0.44	-2.83 – 1.95	0.719	0.02	-2.39 – 2.42	0.990	-0.06	-2.49 – 2.38	0.964	-0.08	-2.53 – 2.37	0.949
Reminder Group x Bot Concern Others				-0.36	-2.87 – 2.16	0.782	-0.62	-3.14 – 1.89	0.627	-0.46	-3.00 – 2.08	0.723	-0.91	-3.51 – 1.68	0.489
Aid Group x Bot Concern Others				0.49	-2.02 – 2.99	0.703	0.33	-2.17 – 2.84	0.793	0.41	-2.12 – 2.93	0.751	0.45	-2.13 – 3.03	0.734
Reminder Group x Bot Concern Future				0.06	-1.08 – 1.19	0.921	-0.03	-1.17 – 1.11	0.962	-0.03	-1.18 – 1.11	0.953	-0.04	-1.18 – 1.11	0.952
Aid Group x Bot Concern Future				-0.99	-2.19 – 0.21	0.105	-1.13	-2.33 – 0.07	0.065	-1.12	-2.32 – 0.08	0.068	-1.15	-2.35 – 0.05	0.061
Social Media Experience							0.07	0.02 – 0.12	**0.006**	0.07	0.02 – 0.12	**0.006**	0.07	0.02 – 0.12	**0.003**
Reminder Group x Social Media Experience							-0.04	-0.11 – 0.04	0.338	-0.04	-0.11 – 0.03	0.312	-0.04	-0.11 – 0.03	0.263
Aid Group x Social Media Experience							-0.07	-0.14 – -0.00	**0.047**	-0.07	-0.14 – -0.00	**0.044**	-0.07	-0.14 – -0.00	**0.042**
Analytical Reasoning (CRT)										0.06	-0.66 – 0.78	0.874	0.14	-0.59 – 0.86	0.708
Reminder Group x CRT										-0.44	-1.47 – 0.58	0.398	-0.47	-1.50 – 0.56	0.373
Aid Group x CRT										-0.21	-1.20 – 0.78	0.677	-0.27	-1.26 – 0.73	0.597
Political View (PV)													0.54	-0.13 – 1.21	0.112
Reminder Group x PV													-0.93	-1.88 – 0.01	0.053
Aid Group x PV													-0.24	-1.19 – 0.72	0.628
Observations	906			906			906			906			906		
R^2 / R^2 adjusted	0.031 / 0.018			0.096 / 0.074			0.105 / 0.081			0.107 / 0.079			0.111 / 0.081		

Model 1 includes participants' conditions, sensitivity, criterion, and related interactions. Those in the control and reminder groups were willing to pay significantly more than those in the aid group, with the control group in between, with means of \$4.91, \$5.92, and \$4.13, respectively. Participants with greater sensitivity and stricter criteria (before treating a persona as a human) were willing to pay more. The effects of sensitivity and criterion varied by test condition in ways that were sustained as other predictors were added to subsequent models. Figures 12 and 13 illustrate these interactions using Model 5 results.

Figure 12

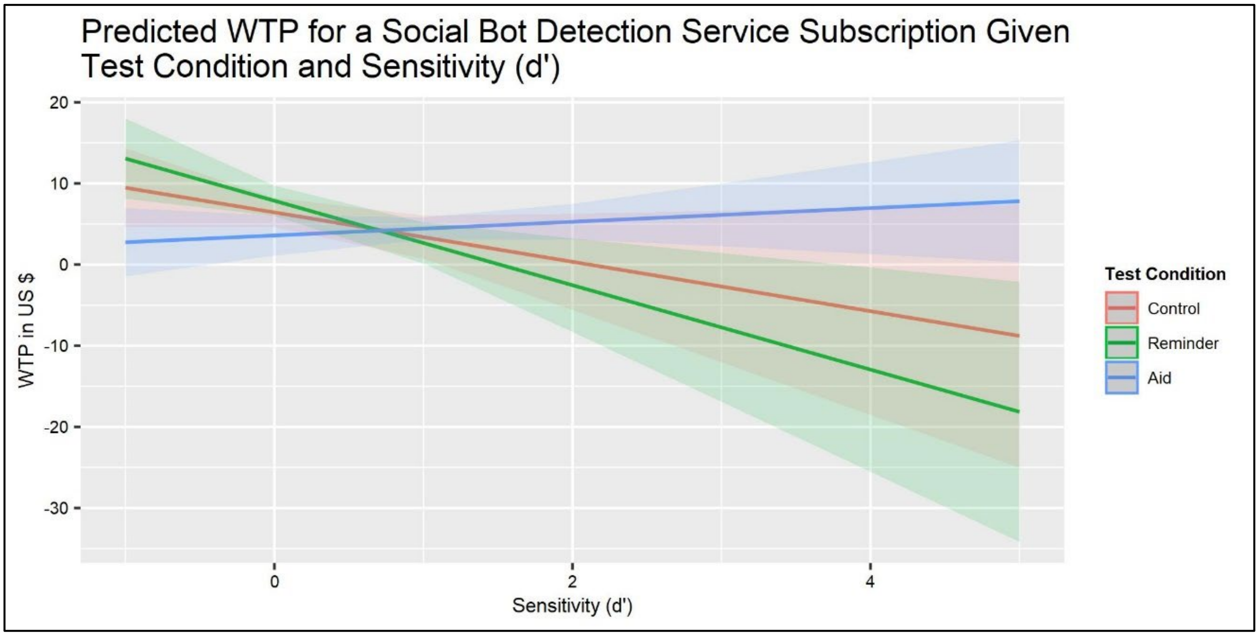
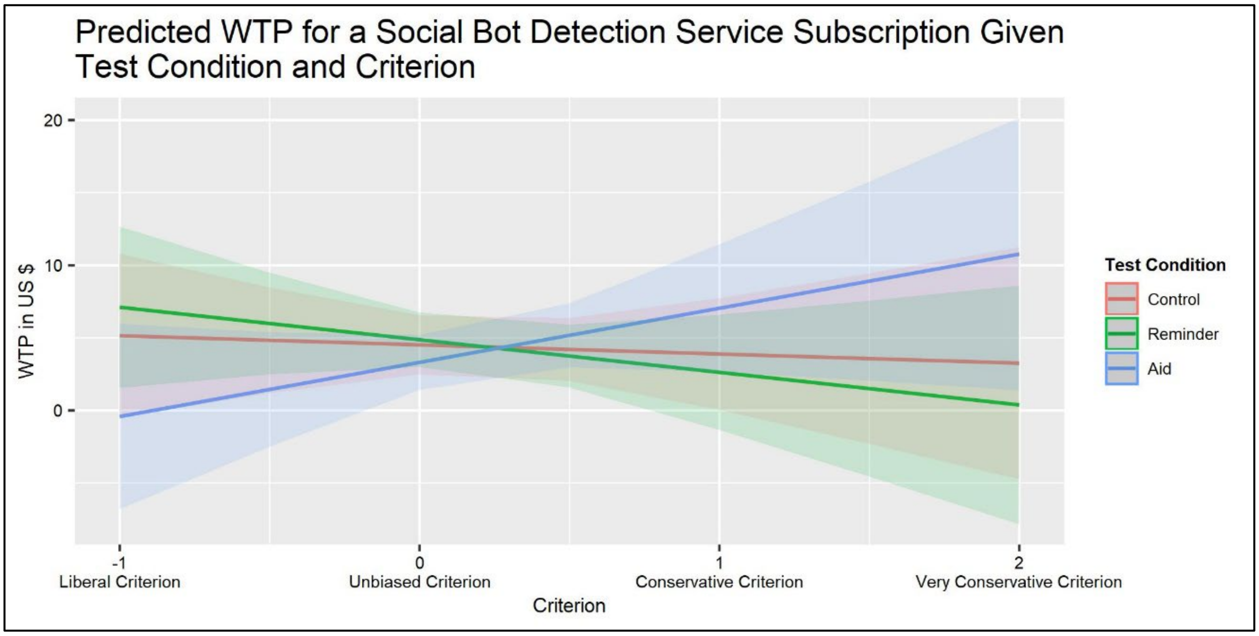


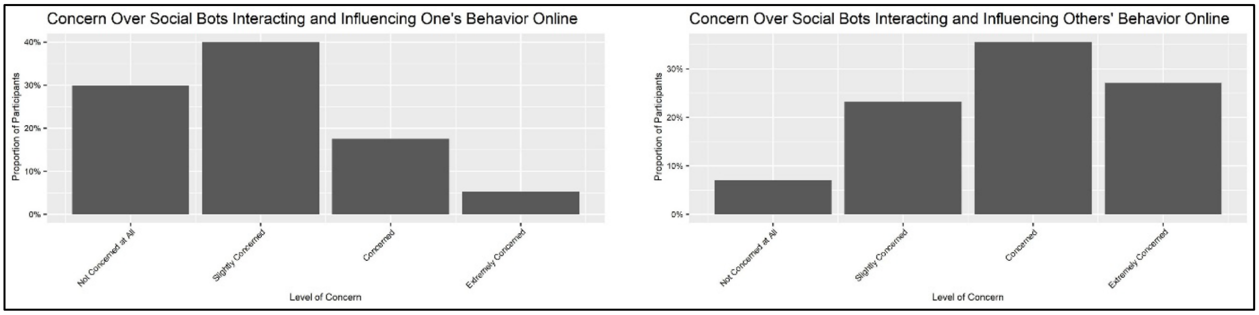
Figure 13



Predicted WTP for a Social Bot Detection Service Subscription Given Test Condition and Criterion

On average, participants reported significantly greater concern about social bots' threat to others than themselves (means=0.98, 1.89, respectively, on a 4-point scale; (p < 0.001). Figure 14 shows the differences.

Figure 14

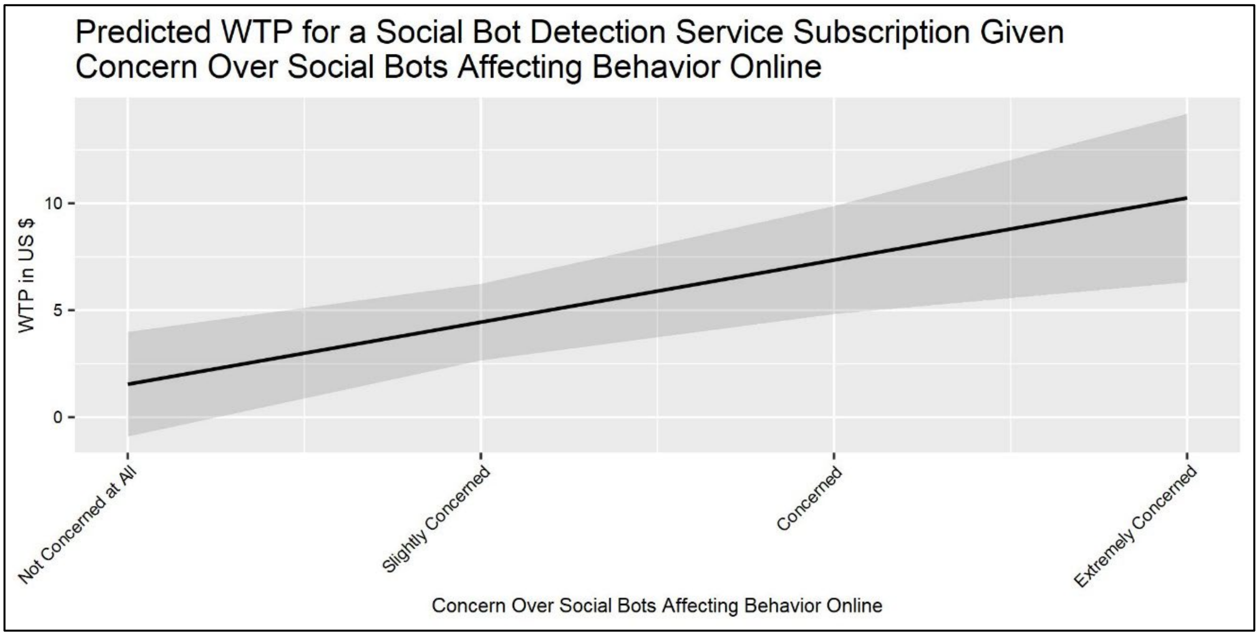


Concern Over Social Bots Interacting and Influencing One's Behavior Online

Concern Over Social Bots Interacting and Influencing Others' Behavior Online

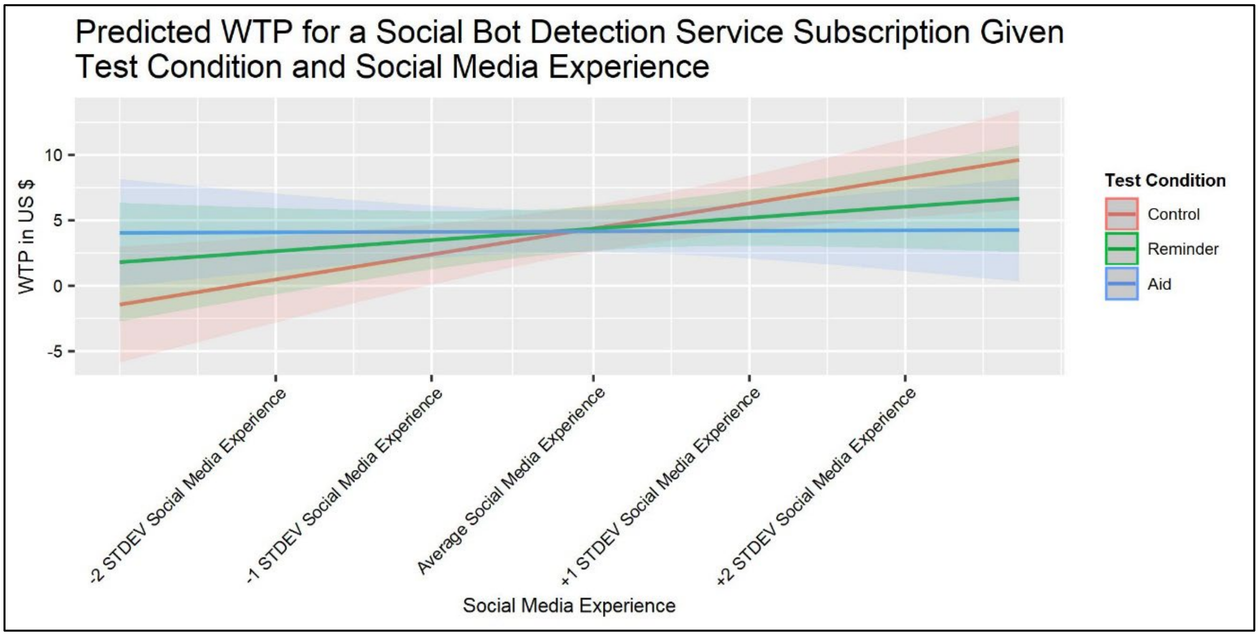
Model 2 in Table 3 adds these concern responses. Participants who reported more significant concern for themselves were willing to pay more for an automated bot indicator, regardless of their condition, sensitivity, or criterion. Figure 15 shows these results. Concern for others was unrelated to a willingness to pay to protect themselves. We also asked when social bots would cause harm to society. Most participants thought that it was happening already (see Appendix E). Given the lack of variability in these responses, we did not conduct exploratory analyses with this variable.

Figure 15



Model 3 adds social media experience, finding that participants who reported having more were willing to pay more for the protection. There was an interaction with the condition, as depicted in Figure 16. Willingness to pay increased with social media experience for participants in the control condition, but not the aid condition – for whom using the aid might have erased any, perhaps illusory, the effect of social media experience. Models 4 and 5 added CRT and PV, finding that neither added predictive value.

Figure 16



Discussion

How decision-makers respond to advice from a human or algorithm depends on how much they trust it and how well they can use it (Hoff & Bashir, 2015; Glikson & Woolley, 2020, Lee & See, 2004; Pavlou, 2003). The current study examined how people use advice from a social bot detection algorithm. Unlike many studies, we found that people made good use of the algorithm, increasing their social bot detection sensitivity and shifting their decision criterion to a more cautious one, seemingly reflecting a better understanding of the threat posed by social bots.

Participants in the control condition had modest sensitivity to the bot signals, replicating Study 1 and supporting H1. The reminder had no effect, contrary to exploratory H4. The aid, showing participants the bot indicator score, significantly increased their sensitivity, supporting H2. Consistent with H3, sensitivity was greatest with the most certain bot indicator scores, closest to 0 and 1. Participants in the control condition had decision criteria that assumed that personas were humans, replicating Study 1 and supporting H5. Participants in the aid condition were more cautious about treating personas as humans; participants in the reminder condition were no different, partially supporting H6. Thus, the decision aid improved performance (sensitivity) and induced greater caution (criterion). The reminder, repeated with each persona, had no effect.

As in Study 1, we observed myside bias, with participants being less ready to treat a persona as a human when it reflected a differing political view, supporting H7. That trend was the same for the three conditions. As in Study 1 (and predicted in H7b), this shift differed for liberal and conservative participants, with liberals' thresholds for responding 'bot' when viewing conservatives decreasing to a greater extent than conservatives' thresholds when considering liberals. Myside bias appeared to dampen sensitivity for conservatives looking at conservatives, similar to the effects found in Kenny et al. (2022, in press), but not for liberals judging liberal personas.

Cognitive Reflection Test (CRT) scores were unrelated to any dependent measure, in any condition, except for one three-way interaction. As seen in Figure 9, in the aid and control conditions, but not the reminder, participants with higher analytical reasoning had more lenient criteria when there were political differences. Supporting only a portion of H11, for participants with and without aid, higher analytical reasoning decreased one's threshold for responding 'bot' when judging a persona of opposing views. However, this amplification of myside bias by analytical reasoning did not materialize for participants provided a reminder to look for bot cues.

Sensitivity was unrelated to social media experience or stimulus presentation order, in any condition, contrary to Study 1 and H11, and H12a. Higher task engagement was associated with less lenient criteria and greater sensitivity, consistent with H12b but not H12c.

Participants rarely were willing to retweet content, even if they agreed with it, especially if they believed it came from a social bot or a human with different political views. The probability of retweeting increased with their confidence in the persona being a human. These findings, along with the strong evidence of myside bias, are consistent with the large body of research into online echo chambers, in which people seek to interact with like-minded people (Garrett, 2009; Flaxman, Goel, & Rao, 2016), on topics such as political ideology (Colleoni, Rozza, & Arvidsson, 2014; Barberá, Jost, Nagler, Tucker, & Bonneau, 2015) climate change (Tyagi, Babcock, Carley, & Sicker, 2020), gun control (Cinelli, Morales, Galeazzi, Quattrociocchi, & Starnini, 2021). and public health (Havey, 2020).

The SDT methodology found that conservatives were more willing than liberals to share content from bot personas if they agreed with the message content. That pattern is consistent with the finding that conservatives take more active roles online (Zhang, Shah, Pevehouse, &

Valenzuela, 2022) and are more willing to promote partisan views strategically, regardless of the source (Freelon, Marwick, & Kreiss, 2020).

Continuing the pattern from Study 1, self-reported social media experience had weak, inconsistent relationships with any dependent measure. If Twitter users rely on their general experience to guide their behavior, they may be misguided – unlike the specific experience of using the social bot indicator aid.

Implications

The aid showed a bot indicator score for each persona, thereby providing implicit feedback regarding the accuracy of participants' intuitive impressions. It led to higher sensitivity, indicating that participants were willing and able to use this form of algorithmic decision support. Those who used the aid were willing to pay more for a social bot detection service subscription, more so if they had cautious criteria (in terms of accepting personas as humans). Nonetheless, they still said bot and human equally often, indicating indifference to the two possible misidentifications in a sample where bot and human personas were equally common.

Participants in the control and reminder conditions demonstrated some sensitivity. They performed well without feedback, indicating some intuitive understanding of personas. Unlike the aid condition, those who performed better were less willing to pay for assistance. Those who assumed that personas were humans regardless of bot signal strength were also less willing to pay more for aid.

The present study replicated and extended the novel reflection of myside bias observed in Study 1. Participants had much higher thresholds for accepting a persona as a human when their political identities differed. – in an experimental setting designed to reflect the online world where polarization occurs (Nikolov, Flammini, & Menczer, 2021). Unfortunately, while the bot detection aid improved sensitivity, it did not reduce myside bias.

Limitations

While our tasks sought to replicate a typical Twitter environment, like any experimental setting, it was somewhat unnatural. One difference from everyday Twitter use is that we instructed participants to look for social bots, which might have enhanced their sensitivity and induced more cautious criteria. The real world has distracted users facing ever-increasing demands on their attention (Qiu, Oliveira, Sahami Shirazi, Flammini, & Menczer, 2017). We would, however, expect the effects of the experimental manipulations (reminder, aid) to be similar.

We also limited our stimuli to a pinned or last tweet within a profile, denying participants the chance to examine more of a persona's tweets, potentially reducing their performance. Thus, we may have underestimated sensitivity to the extent that Twitter users investigate personas more thoroughly in real-world settings.

The limited predictive value of social media experience and analytical reasoning ability should reduce confidence in the seemingly plausible relationships posited in our hypotheses. However, it may also mean that our survey instruments failed to capture the constructs in these hypotheses. The social media experience survey, adapted from Hou (2017), asks about activity levels and depth of engagement for a much broader range of activities than Twitter. The Cognitive Reflection Test asks much more general questions. In that light, these results provide insight into those measures' range of applicability. Our reminder condition might have been weaker than the kind of accuracy nudge that some have proposed (e.g., Pennycook & Rand, 2021).

Conclusion

This study provides evidence for an intervention that could help reduce social bot deception and an example of a usable algorithm-based decision aid. Participants who responded 'bot' were far less willing to share personas' content, suggesting that better discrimination could help reduce the spread of false and misleading information. However, the aid still left imperfect sensitivity and substantial myside bias. Moreover, our aid is only relevant to users who have access to algorithm support. Study 3 addresses situations where that support is lacking.

Key Points

- We evaluated performance in distinguishing Twitter personas produced by humans or social bots, using an AI-based aid, a reminder to look for cues, and no aid (control).
- We found that training increased sensitivity and shifted participants' criterion to be warier of bots.
- We found evidence of myside bias, with individuals more likely to accept personas as human if they have similar political views.
- Individuals were willing to pay more for bot detection aid to the extent that they showed greater sensitivity to bots, were more cautious about treating personas as bots, expressed greater concern about bots, and spent more time on social media.

Ch 4. Training Bot Detection

Social bot detection tools have been developed to help users identify bots. Study 2 found that individuals could use a bot indicator score to improve (although not perfect) their judgments of Twitter personas. Participants who received the aid had greater social bot detection sensitivity and more balanced decision criteria than participants who received no aid or just a reminder to look for bot cues, which had little effect.

Study 3 replicates Study 2's evaluation of the social bot indicator aid, testing the robustness of its findings. Study 3 also addresses the situation facing individuals without access to such assistance, as could happen for various reasons. Social media platforms may resist employing bot indicators that could reveal their algorithms, add costs that cannot be passed on to users, or cause embarrassment and distrust (e.g., by mislabeling personas) (Constantin, 2022; Jagielski, Oprea, Biggio, Liu, Nita-Rotaru, & Li, 2018). Creating valid indicator scores requires large data sets, which may not be available for new personas, perhaps launched for short-term gains.

Even when users have access, they may not trust such automated tools. Although algorithms can help detect and target diseases, their adoption has been slow (Dietvorst, Simmons, & Massey, 2015; Buck, Doctor, Hennrich, Jöhnk, & Eymann, 2022; Gaube, Suresh, Raue, Merritt, Berkowitz, Lermer,... & Ghassemi, 2021). Medical algothrims can improve diagnosis and treatment, yet patients prefer human medical advice (Yokoi, Eguchi, Fujita, & Nakayachi, 2021; Longoni, Bonezzi, & Morewedge, 2019). Automatic algorithms can improve hiring practices by reducing biases and filtering for preferred job skills, yet hiring practitioners prefer human input (Tambe, Cappelli, & Yakubovich, 2019; Diab, Pui, Yankelevich, & Highhouse, 2011). Although computational models may seem like natural ways to manage the complexity of investing, many people prefer advice from human experts (Önkal, Goodwin, Thomson, Gönül, & Pollock, 2009). Thus, even when a valuable bot detection aid is available, people might not use it (Holzinger, Kieseberg, Weippl, & Tjoa, 2018), a reluctance that may have limited improvement in Study 2.

A further reason to avoid automated aids is fear of complacency and dependence (Bahner, Hüper, & Manzey, 2008; Dwivedi, Hughes, Ismagilova, Aarts, Coombs, Crick,... & Williams, 2021). Banker and Khatani (2019) found that consumers uncritically accepted recommendations from an inferior algorithm, overestimating its value. Likewise, the regular use of GPS has eroded the spatial memory needed for self-guided navigation (Dahmani & Bohbot, 2020). Should social media users become overly reliant upon bot indicator aids, they may gradually lose their ability to detect social bots without assistance.

We develop and evaluate an alternative strategy that, if successful, could address these concerns: training users in heuristic rules for identifying bots. Those heuristics would not depend on social media platforms to provide (or allow) automated bot detectors. They would always be available, even for new personas with too little data for algorithms to analyze. Moreover, they would require ongoing user engagement, reducing the risk of complacency.

As with any intervention, success depends on the quality of execution. The heuristics must be valid predictors of whether personas are social bots. The training must convey these heuristics in intuitively meaningful terms. The users must have the trust and skills needed to apply them. We will answer these empirical questions using the methodology from Study 2. We will see whether people exposed to the training have greater sensitivity and more balanced decision criteria than people who are not. We will also compare them with people who receive the more intensive (and expensive) bot indicator intervention. We will examine how intervention effects vary with the individual attributes of social media experience, analytical reasoning, and myside bias.

Training Social Bot Detection

Judging whether a Twitter persona is a social bot is a classification task with consequences. Failing to identify social bots may lead to deception, endorsement of false narratives, or amplification of harmful information operations. Mislabeling humans as social bots, on the other hand, may isolate users from credible perspectives and valuable social networks. Therefore, helping to improve the accuracy of social media users' social bot detection can improve individuals' online experiences and potentially decrease information pollution proliferated by social bots. They would promote active learning, as users get feedback on the accuracy of their heuristic predictions.

Professionals refine their heuristics through experience, accompanied by quality feedback (Kahneman & Klein, 2009). However, experience alone does not guarantee quality judgments

(Ericsson, 2017; Hanushek & Rivkin, 2010) unless, as Simon and Chase (1973) proposed, it translates into the pattern recognition essential to heuristic search and inference (Ericsson, 2017). Similar to how physicians can learn to refine their heuristics for trauma triage (Mohan et al., 2017; Mohan et al., 2014), social media users may be able to refine their heuristics for social bot detection. That training sought heuristic rules that (a) had diagnostic value (for triage decisions), (b) were not intuitively obvious (hence needed training), and (c) could be easily integrated with physicians' natural ways of thinking (or mental models). We sought to meet the same conditions for improving social media users' heuristics for detecting social bots.

Determining Normative Expectations of Social Bot Detection Training

Thus, our first step was to identify heuristic rules with a diagnostic value that social media users could apply. Machine learning algorithms claim greater than 90% accuracy in detecting social bots. However, these algorithms have access to user data and network analytics. Moreover, they use both persona-defined information, which can be easily manipulated, and platform-managed information, which is much harder to exploit. Previous exploratory analysis of participants' ratings of the utility of profile features in informing their bot detection judgments led us to speculate they expected to see greater bot signals amongst persona-defined information (see Appendix G). We sought cues in publicly visible Twitter profiles with sufficient bot signal strength to have the potential for training and enough intuitive resonance to become user heuristics.

Following Mohan et al. (2014, 2017), we sought diagnostic cues that reflected a common strategy, allowing users to create a single heuristic (rather than master a set of unrelated rules). The common strategy is looking for personas that are inordinately focused on reaching as many people as possible as fast as possible. We conjectured that two cues would meet these criteria: (a) a large number of Tweets relative to the age of the account, reflecting an attempt to amplify narratives quickly; and (b) a large number of followed personas, reflecting an attempt to expand a social network, hoping to enlist followers.

We tested this conjecture by evaluating the predictive value of cues that represent this strategy and are readily found on a Twitter persona profile page. Appendix H describes these experiments, which use the ability to predict Bot Hunter's Tier 1 results as their criterion. The first experiment trained four machine learning algorithms using four visible (Tier 1) platform-managed features selected because they suggest an amplification strategy: account age, total number of tweets, number of following, and number of followers. We found that the best algorithm using these four features had an accuracy of 85.8%. The second experiment applied the same four machine learning algorithms to the visible features not used in the first experiment. It found that they had much less predictive value, with the best model having predictive accuracy of only 68.5%.

We validated the results of the first experiment, which used an extensive training set, for our stimuli by assessing the accuracy of the final trained algorithms on the 60 experimental stimuli. The best algorithm was 90% accurate, giving us confidence that these features had sufficient signal strength to inform participants' judgments and support our training – if people will and can use them.

Predicted Relationships

This study compared participants randomly assigned to one of three groups, each of which judged the same 60 Twitter personas by examining their Twitter profiles. All three groups first viewed a brief introductory video. The *training* group then viewed an instructional video on social bot detection. The *aid* group received a bot indicator score near each persona's name, as in Study 2, and no additional video. The *control* group went directly to the persona evaluation task. We group our hypotheses by task and participant characteristics.

As before, we have two primary dependent measures:

Sensitivity is measured as the change in participants' probability of saying "bot" as the bot indicator score increases.

Criterion is participants' tendency to identify stimuli as bots or humans. If bot and human personas are equally likely (as is the case here), that tendency indicates participants' relative aversion to mistaking a bot for a human and vice versa. Studies 1 and 2 found that participants require more evidence to call a stimulus a bot than a human, reflected in negative criterion scores.

Task (Experimental Condition)

Sensitivity

(H1) Based on the sensitivity observed in Studies 1and 2, we expect the probability of participants responding 'bot' to increase as the bot signal increases for all three conditions.

(H2) As in Study 2, we expect participants in the aid condition to perform better than participants in the control condition, reflecting their ability to use the bot indicator score.

(H3) We expected participants in the training condition to use at least some of the information in the training, improving their sensitivity relative to the control group.

Criterion

(H4) Based on Studies 1 and 2, we expected control group participants' criteria to be negative (in our scoring), reflecting greater aversion to mistaking a human for a bot than vice versa.

(H5) Based on Study 2, we expected participants in the aid condition to show less aversion to mistaking a human for a bot, reflecting in less negative criteria and a higher proportion of "bot" responses than in the control condition.

(H6) The training group learns about social bot developers' motives. If that knowledge makes the consequences of successful deception seem worse, participants may be more averse to mistaking a bot for a human, shifting their criterion.

Controls

As in Studies 1 and 2, we included two control variables in our prediction models with the same predictions.

(H7) Participants' sensitivity will decrease with stimulus presentation order, reflecting fatigue (Parasuraman & Davies, 1977).

(H8) Participants with higher task engagement scores will have greater sensitivity (Kenny et al., 2022; Matthews, Warm, Reinerman, Langheim & Saxby, 2010; Downs, Holbrook, Sheng & Cranor, 2010; Dewitt, Fischhoff, Davis & Broomell, 2015; Macmillan & Creelman, 2004) and show greater improvements in sensitivity with training.

(H9) Neither control variable will be related to participants' decision criteria.

Participant Characteristics

Social Media Experience. Studies 1 and 2 found that social media experience was either unrelated or negatively related to participants' sensitivity and unrelated to their criteria. Although studies of expertise (Ericsson & Pool, 2016) have found that experience can increase sensitivity, our investigations led us to predict no relationship.

(H10) Social media experience will be unrelated to sensitivity and criteria in all three conditions.

Analytical Reasoning Ability. Prior research led us to expect participants with higher CRT scores to have greater sensitivity and be less hesitant to respond 'bot,' as reflected in their criteria. Study 1 affirmed that prediction. Study 2 did not. Given prior research, we will retain our initial prediction.

(H11) Participants with higher CRT scores will have (H1a) greater sensitivity and (H11b) criteria more averse to mistaking bots for humans.

Myside Bias. Studies 1 and 2 found evidence of myside bias affecting participants' criteria and sensitivity. Both liberals and conservatives set a higher threshold for classifying a stimulus as a bot when it agreed with their political views. Exploratory analyses found that the tendency was stronger for liberals than for conservatives, whose judgments of conservative persona demonstrated little sensitivity to bot signals. As the training does not address myside bias, we expect myside bias in all test conditions, as before.

(H12) Myside bias will appear in all conditions and repeat the differential effects with political liberals and conservatives observed in Studies 1 and 2.

Study 2 revealed greater myside bias among participants with higher CRT scores, as though they used that ability to defend existing positions.

(H13) Participants with higher CRT scores will display greater myside bias.

Methods

Sample

We collected data in March 2022. We recruited participants (N = 924) using Prolific (Palan & Schitter, 2018) and paid them $8 for approximately 30 minutes of work. Participation was limited to US citizens and native English speakers. Informed consent was obtained. The research followed the American Psychological Association Code of Ethics and was approved by the Carnegie Mellon University Institutional Review Board under protocol # IRB00000603.

Design

Participants judged the same 60 Twitter personas, using the same protocol as Study 2. Participants examined a public Twitter persona profile in each trial and indicated whether they believed a bot or human-produced it. They then rated their confidence in that choice, using a slide bar on a scale anchored at 50% (completely uncertain) and 100% (certain). Finally, they indicated whether they would share content from the persona by retweeting it if they agreed with its content.

We assigned participants randomly to three experimental groups, which received the bot detector aid (aid), the training video (training), or nothing else (control). As in Study 2, the aid group received the bot indicator score near the persona's name.

Training Protocol

Before completing any SDT trials, all participants watched a 40-second video that defined social bots and highlighted their increasing presence on social media. Participants in the training condition then watched two additional videos. The first, lasting approximately 90 seconds, explained the importance of identifying and then ignoring information that the bot creator wants them to see, knowing that bot creators' central objective is getting a message to as many people as possible as quickly as possible. We instructed participants to examine:

(1) the persona's total number of tweets, and

(2) the account's age.

They were then told, "if an account is relatively new, yet has produced a high volume of Tweets – they are likely seeking to influence people – and may be a social bot using automation to do so."

Next, we instructed them to examine the size of the persona's social network in terms of

(3) how many people they are following, and

(4) how many people were following them.

They were then told, "If the persona's social network appears disproportionally large, they may be part of a bot network. Or, if the ratio of following to followers is high, they may also be a social bot."

Training group participants watched a final two-minute video that led them through four examples demonstrating the decision rules. We asked two comprehension questions following the training to keep participants engaged. Almost all answered both questions correctly.

After completing the SDT trials, all participants completed the same demographic survey and questions as in Study 2, assessing social media experience, CRT performance, and political views. For exploratory purposes, we elicited their willingness to pay for "an automated social bot detection service" and their concern over social bots as a threat to themselves and others. As an input to future research, we also tested an exploratory measure that elicited participants' willingness to use a service that delayed loading suspicious profiles. We did not analyze these responses.

Analysis Plan

Our primary dependent variable is the probability of responding 'bot' to a Twitter persona. The analysis plan employs general linear mixed-effects regression to estimate the predictive value of the fixed effects of experimental conditions (control, aid, training), stimuli attributes (bot indicator score, political tone), and individual characteristics (social media experience, cognitive reflection ability, and political view) for responding 'bot.' We also calculated a *political disagreement* score, equal to the difference between our ratings of each persona's political tone and each participant's political self-rating. We included stimulus presentation order and participant task engagement as control variables.

All models employed a Probit link function. To predict the probability of responding 'bot,' z-scores for an observation are converted to probabilities by taking the inverse of Phi. Fixed effects for each model coefficient were estimated with a multivariate Gaussian distribution, using the arm package in R (Gelman, Su, Yajima, Hill, Pittau, Kerman & Dorie, 2016).

We followed analogous analytical strategies for the other dependent variables (e.g., willingness to retweet).

Results

Sample Demographics

Nine hundred and eighty-two participants completed the study. Following our preregistered procedure, we excluded two participants for failing to follow Prolific's verification procedures, two for completing the task too quickly (less than six minutes), and two for taking too long (more than three hours).

The analyzed sample included 976 participants, 433 males, 531 females, 10 non-binary, and 2 preferred to say; age ranged from 18 to 84 years old (median = 36; mean = 38). Seven hundred eighty-three reported being White, 55 Hispanic or Latino, 50 Black or African American, 5 Native American, 64 Asian or Pacific Islander, and 19 Other. Nine reported less than high school education, 403 a high school degree or equivalent, 429 a bachelor's degree, 107 a master's degree, and 28 a doctorate. Four hundred sixty-five reported being fully employed, 123 employed part-time, 72 unemployed and looking, 59 unemployed and not looking, 89 students, 54 retired, 94 self-employed, and 20 unable to work. Three hundred eighty-eight reported being married, 12 widowed, 78 divorced, 8 separated, and 490 never married. Annual incomes were roughly normally distributed, over 8 categories ranging from "less than 10K" to "over 150K," with the median between 25K-50K.

Criterion, Bot Indicator, and Controls

As in Studies 1 and 2, the estimated SDT parameters assumed that the signal and noise distributions were Gaussian with equal variance (Lynn & Barrett, 2014). A log-linear correction added 0.5 to the number of hits and false alarms and 1 to the number of signals (bot personas) or noise (human personas) to correct for participants who identified all stimuli correctly or incorrectly, thereby producing hit (H) or false alarm (FA) rates of 0 or 1 (Hautus,199). Thus, d' and c were calculated using: $H = (hits + 0.5)/(signals + 1)$

$$FA = \ (false\ alarms\ 0.5)/(noise\ +\ 1)$$

$$d'\ (sensitivity) =\ z(H) - z(FA)$$

$$c\ (criterion)\ = - 0.5[z(H)\ +\ z(FA)]$$

Sensitivity. Figure 1 displays sensitivity (d') in the three test conditions. Participants in all three conditions showed sensitivity to the bot indicator score, as evidenced by the upward trend of the probability curves (affirming H1). However, the aid and training conditions showed greater sensitivity than the control condition (affirming H2 and H3).

Table 1 summarizes probit regressions predicting the probability of judging a persona to be a bot. Model 1 addresses the primary patterns in Figure 1, controlling for Task Engagement (TE) and Serial Presentation Order (SPO). The positive coefficient for bot indicator score (BI) in these interactions reflects the overall upward trend. The significant interactions with BI in Model 1 reflect the greater sensitivity of the Training (BIxTG) and Aid (BIxAG) groups.

Figure 1

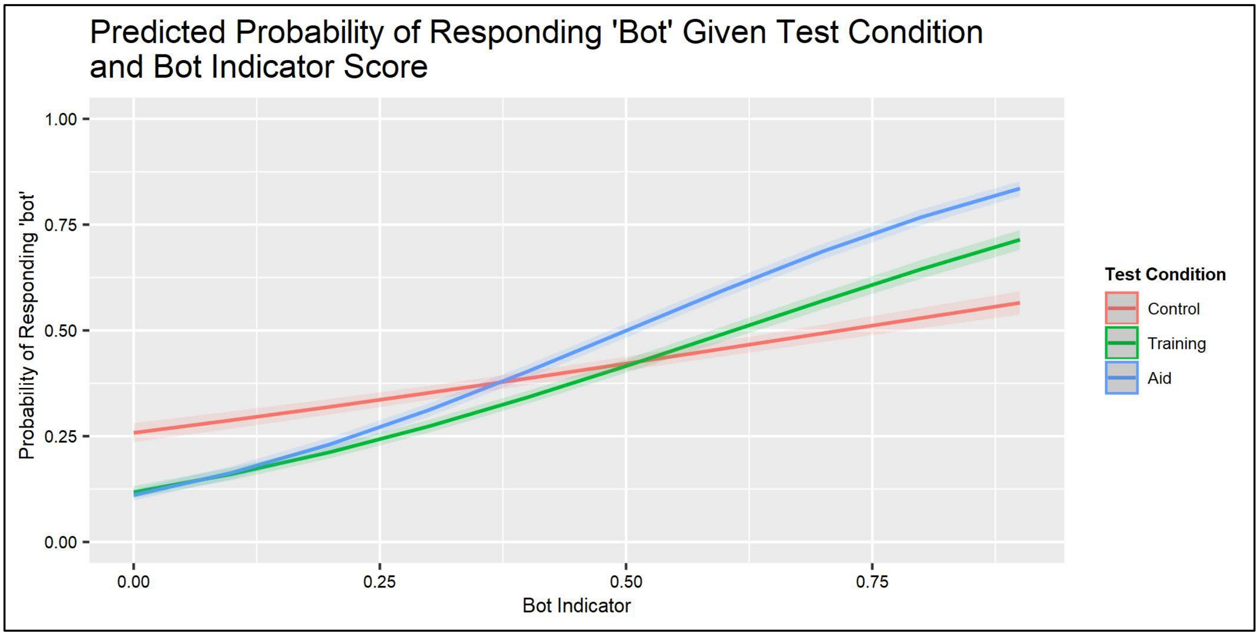
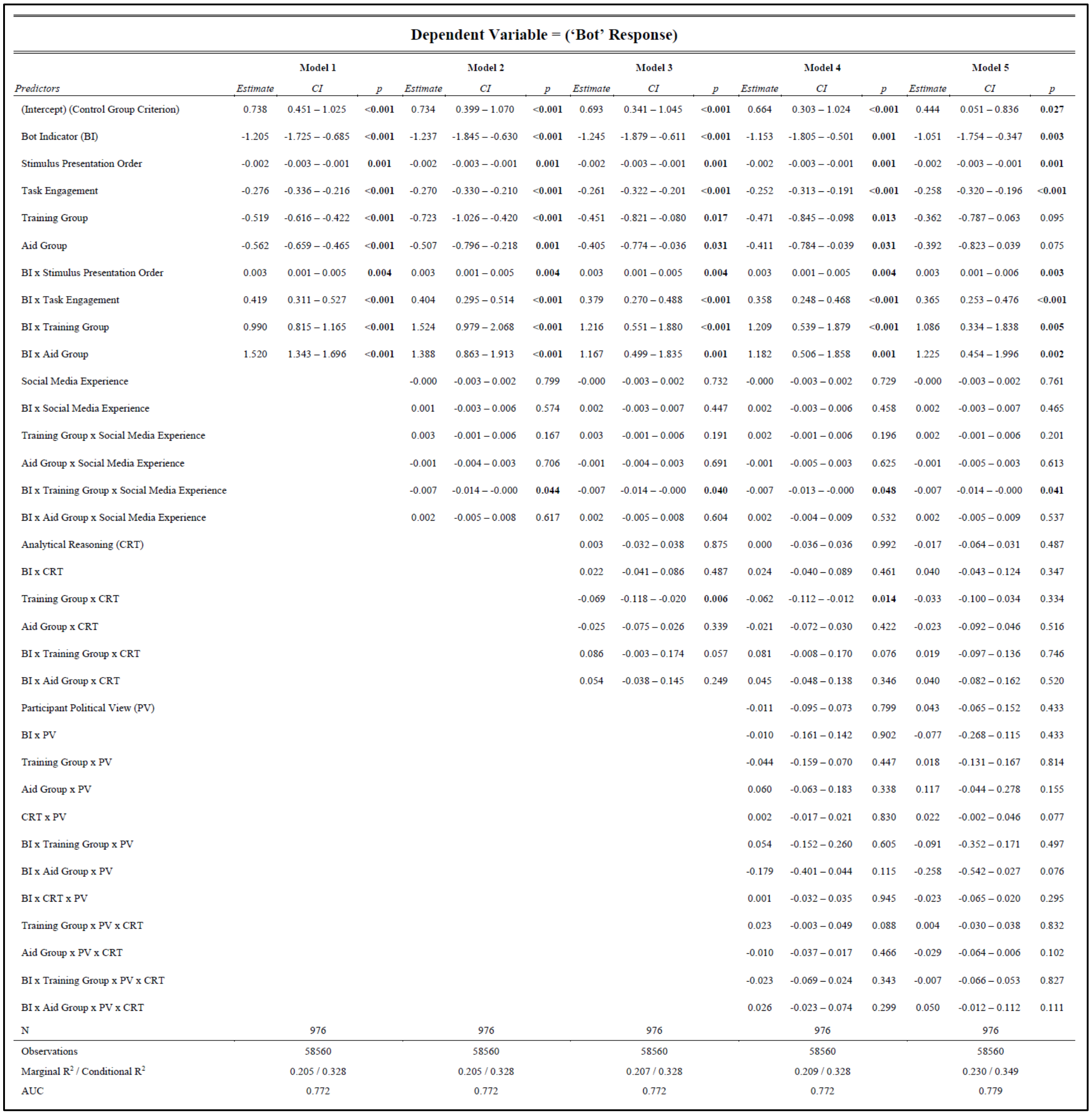


Table 1a

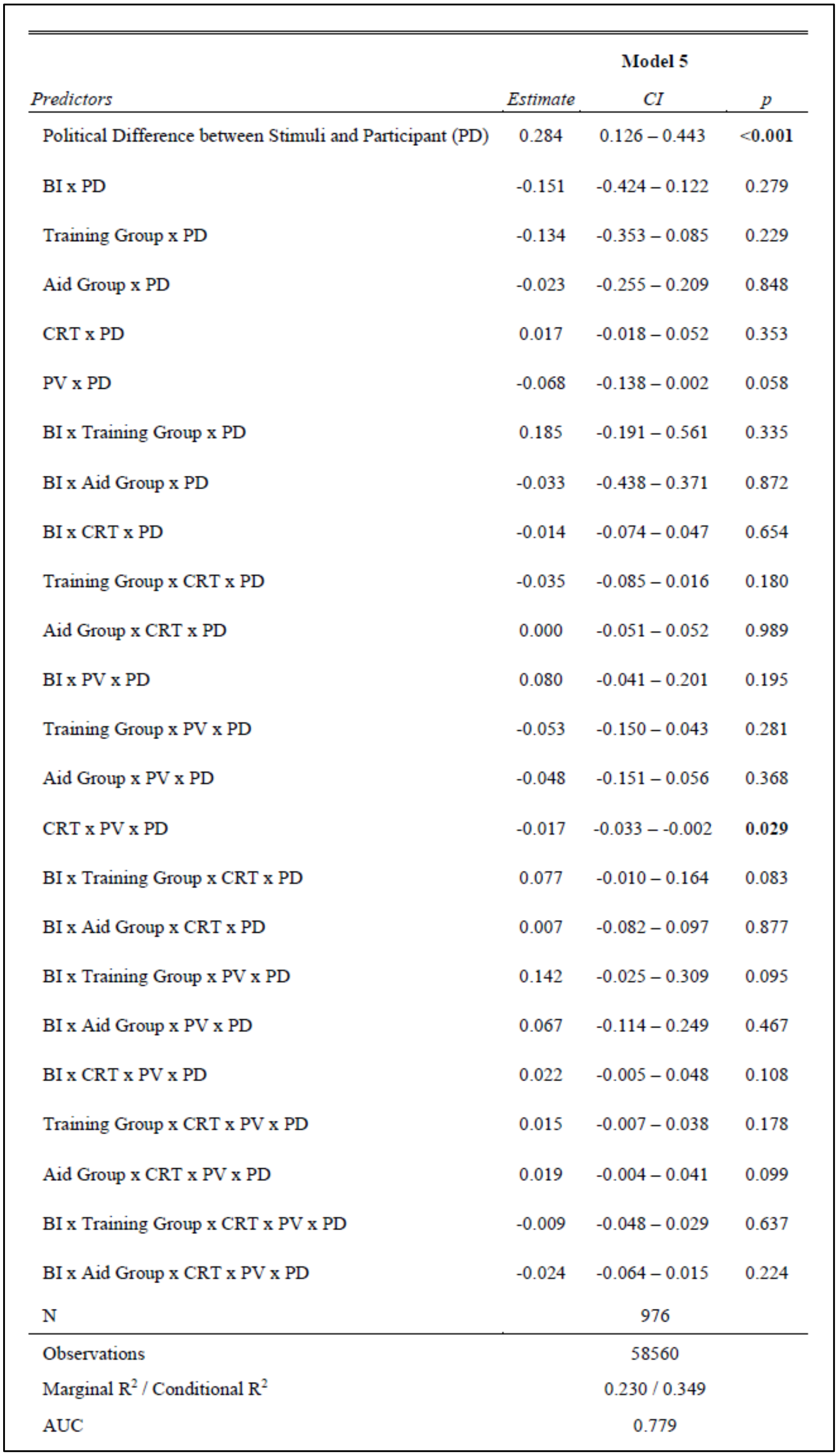
General Linear Mixed Effects Probit Regression Models, Predicting the Probability of Judging a Persona to Be a Bot.



Dependent Variable = ('Bot' Response)

	Model 1			**Model 2**			**Model 3**			**Model 4**			**Model 5**		
Predictors	*Estimate*	*CI*	*p*	*Estimate*	*CI*	*p*	*Estimate*	*CI*	*p*	*Estimate*	*CI*	*p*	*Estimate*	*CI*	*p*
(Intercept) (Control Group Criterion)	0.738	0.451 – 1.025	**<0.001**	0.734	0.399 – 1.070	**<0.001**	0.693	0.341 – 1.045	**<0.001**	0.664	0.303 – 1.024	**<0.001**	0.444	0.051 – 0.836	**0.027**
Bot Indicator (BI)	-1.205	-1.725 – -0.685	**<0.001**	-1.237	-1.845 – -0.630	**<0.001**	-1.245	-1.879 – -0.611	**<0.001**	-1.153	-1.805 – -0.501	**0.001**	-1.051	-1.754 – -0.347	**0.003**
Stimulus Presentation Order	-0.002	-0.003 – -0.001	**0.001**	-0.002	-0.003 – -0.001	**0.001**	-0.002	-0.003 – -0.001	**0.001**	-0.002	-0.003 – -0.001	**0.001**	-0.002	-0.003 – -0.001	**0.001**
Task Engagement	-0.276	-0.336 – -0.216	**<0.001**	-0.270	-0.330 – -0.210	**<0.001**	-0.261	-0.322 – -0.201	**<0.001**	-0.252	-0.313 – -0.191	**<0.001**	-0.258	-0.320 – -0.196	**<0.001**
Training Group	-0.519	-0.616 – -0.422	**<0.001**	-0.723	-1.026 – -0.420	**<0.001**	-0.451	-0.821 – -0.080	**0.017**	-0.471	-0.845 – -0.098	**0.013**	-0.362	-0.787 – 0.063	0.095
Aid Group	-0.562	-0.659 – -0.465	**<0.001**	-0.507	-0.796 – -0.218	**0.001**	-0.405	-0.774 – -0.036	**0.031**	-0.411	-0.784 – -0.039	**0.031**	-0.392	-0.823 – 0.039	0.075
BI x Stimulus Presentation Order	0.003	0.001 – 0.005	**0.004**	0.003	0.001 – 0.005	**0.004**	0.003	0.001 – 0.005	**0.004**	0.003	0.001 – 0.005	**0.004**	0.003	0.001 – 0.006	**0.003**
BI x Task Engagement	0.419	0.311 – 0.527	**<0.001**	0.404	0.295 – 0.514	**<0.001**	0.379	0.270 – 0.488	**<0.001**	0.358	0.248 – 0.468	**<0.001**	0.365	0.253 – 0.476	**<0.001**
BI x Training Group	0.990	0.815 – 1.165	**<0.001**	1.524	0.979 – 2.068	**<0.001**	1.216	0.551 – 1.880	**<0.001**	1.209	0.539 – 1.879	**<0.001**	1.086	0.334 – 1.838	**0.005**
BI x Aid Group	1.520	1.343 – 1.696	**<0.001**	1.388	0.863 – 1.913	**<0.001**	1.167	0.499 – 1.835	**0.001**	1.182	0.506 – 1.858	**0.001**	1.225	0.454 – 1.996	**0.002**
Social Media Experience				-0.000	-0.003 – 0.002	0.799	-0.000	-0.003 – 0.002	0.732	-0.000	-0.003 – 0.002	0.729	-0.000	-0.003 – 0.002	0.761
BI x Social Media Experience				0.001	-0.003 – 0.006	0.574	0.002	-0.003 – 0.007	0.447	0.002	-0.003 – 0.006	0.458	0.002	-0.003 – 0.007	0.465
Training Group x Social Media Experience				0.003	-0.001 – 0.006	0.167	0.003	-0.001 – 0.006	0.191	0.002	-0.001 – 0.006	0.196	0.002	-0.001 – 0.006	0.201
Aid Group x Social Media Experience				-0.001	-0.004 – 0.003	0.706	-0.001	-0.004 – 0.003	0.691	-0.001	-0.005 – 0.003	0.625	-0.001	-0.005 – 0.003	0.613
BI x Training Group x Social Media Experience				-0.007	-0.014 – -0.000	**0.044**	-0.007	-0.014 – -0.000	**0.040**	-0.007	-0.013 – -0.000	**0.048**	-0.007	-0.014 – -0.000	**0.041**
BI x Aid Group x Social Media Experience				0.002	-0.005 – 0.008	0.617	0.002	-0.005 – 0.008	0.604	0.002	-0.004 – 0.009	0.532	0.002	-0.005 – 0.009	0.537
Analytical Reasoning (CRT)							0.003	-0.032 – 0.038	0.875	0.000	-0.036 – 0.036	0.992	-0.017	-0.064 – 0.031	0.487
BI x CRT							0.022	-0.041 – 0.086	0.487	0.024	-0.040 – 0.089	0.461	0.040	-0.043 – 0.124	0.347
Training Group x CRT							-0.069	-0.118 – -0.020	**0.006**	-0.062	-0.112 – -0.012	**0.014**	-0.033	-0.100 – 0.034	0.334
Aid Group x CRT							-0.025	-0.075 – 0.026	0.339	-0.021	-0.072 – 0.030	0.422	-0.023	-0.092 – 0.046	0.516
BI x Training Group x CRT							0.086	-0.003 – 0.174	0.057	0.081	-0.008 – 0.170	0.076	0.019	-0.097 – 0.136	0.746
BI x Aid Group x CRT							0.054	-0.038 – 0.145	0.249	0.045	-0.048 – 0.138	0.346	0.040	-0.082 – 0.162	0.520
Participant Political View (PV)										-0.011	-0.095 – 0.073	0.799	0.043	-0.065 – 0.152	0.433
BI x PV										-0.010	-0.161 – 0.142	0.902	-0.077	-0.268 – 0.115	0.433
Training Group x PV										-0.044	-0.159 – 0.070	0.447	0.018	-0.131 – 0.167	0.814
Aid Group x PV										0.060	-0.063 – 0.183	0.338	0.117	-0.044 – 0.278	0.155
CRT x PV										0.002	-0.017 – 0.021	0.830	0.022	-0.002 – 0.046	0.077
BI x Training Group x PV										0.054	-0.152 – 0.260	0.605	-0.091	-0.352 – 0.171	0.497
BI x Aid Group x PV										-0.179	-0.401 – 0.044	0.115	-0.258	-0.542 – 0.027	0.076
BI x CRT x PV										0.001	-0.032 – 0.035	0.945	-0.023	-0.065 – 0.020	0.295
Training Group x PV x CRT										0.023	-0.003 – 0.049	0.088	0.004	-0.030 – 0.038	0.832
Aid Group x PV x CRT										-0.010	-0.037 – 0.017	0.466	-0.029	-0.064 – 0.006	0.102
BI x Training Group x PV x CRT										-0.023	-0.069 – 0.024	0.343	-0.007	-0.066 – 0.053	0.827
BI x Aid Group x PV x CRT										0.026	-0.023 – 0.074	0.299	0.050	-0.012 – 0.112	0.111
N	976			976			976			976			976		
Observations	58560			58560			58560			58560			58560		
Marginal R^2 / Conditional R^2	0.205 / 0.328			0.205 / 0.328			0.207 / 0.328			0.209 / 0.328			0.230 / 0.349		
AUC	0.772			0.772			0.772			0.772			0.779		

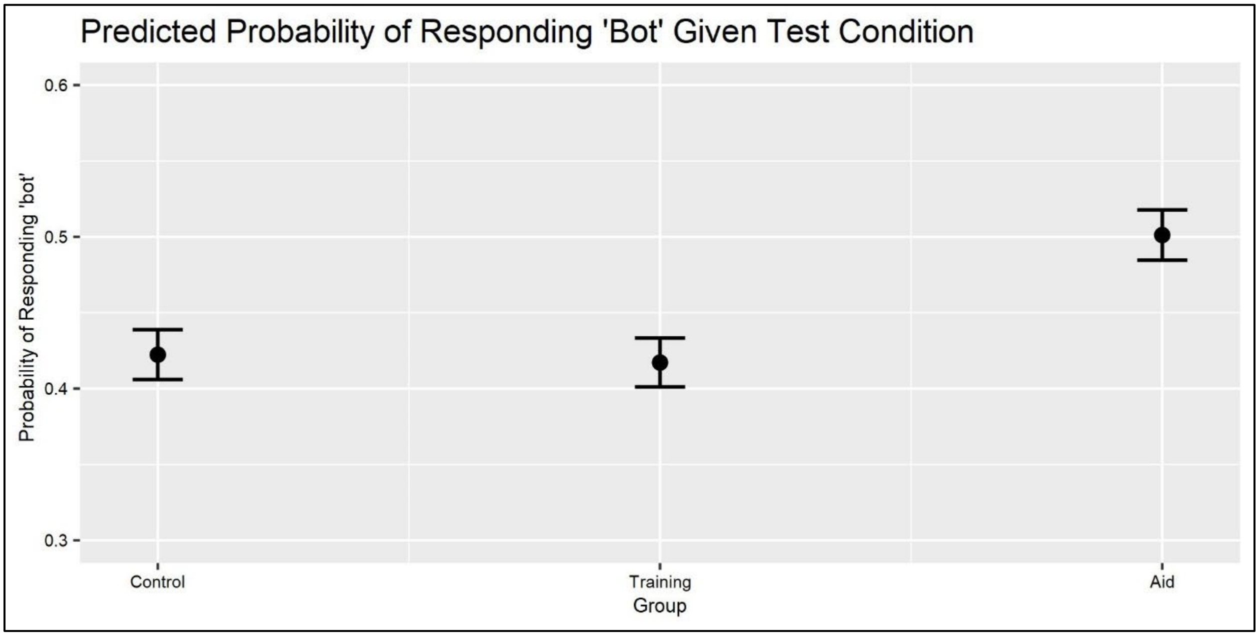
Table 1b



		Model 5	
Predictors	*Estimate*	*CI*	*p*
Political Difference between Stimuli and Participant (PD)	0.284	0.126 – 0.443	**<0.001**
BI x PD	-0.151	-0.424 – 0.122	0.279
Training Group x PD	-0.134	-0.353 – 0.085	0.229
Aid Group x PD	-0.023	-0.255 – 0.209	0.848
CRT x PD	0.017	-0.018 – 0.052	0.353
PV x PD	-0.068	-0.138 – 0.002	0.058
BI x Training Group x PD	0.185	-0.191 – 0.561	0.335
BI x Aid Group x PD	-0.033	-0.438 – 0.371	0.872
BI x CRT x PD	-0.014	-0.074 – 0.047	0.654
Training Group x CRT x PD	-0.035	-0.085 – 0.016	0.180
Aid Group x CRT x PD	0.000	-0.051 – 0.052	0.989
BI x PV x PD	0.080	-0.041 – 0.201	0.195
Training Group x PV x PD	-0.053	-0.150 – 0.043	0.281
Aid Group x PV x PD	-0.048	-0.151 – 0.056	0.368
CRT x PV x PD	-0.017	-0.033 – -0.002	**0.029**
BI x Training Group x CRT x PD	0.077	-0.010 – 0.164	0.083
BI x Aid Group x CRT x PD	0.007	-0.082 – 0.097	0.877
BI x Training Group x PV x PD	0.142	-0.025 – 0.309	0.095
BI x Aid Group x PV x PD	0.067	-0.114 – 0.249	0.467
BI x CRT x PV x PD	0.022	-0.005 – 0.048	0.108
Training Group x CRT x PV x PD	0.015	-0.007 – 0.038	0.178
Aid Group x CRT x PV x PD	0.019	-0.004 – 0.041	0.099
BI x Training Group x CRT x PV x PD	-0.009	-0.048 – 0.029	0.637
BI x Aid Group x CRT x PV x PD	-0.024	-0.064 – 0.015	0.224
N		976	
Observations		58560	
Marginal R^2 / Conditional R^2		0.230 / 0.349	
AUC		0.779	

Criterion. Figure 2 displays the distributions of individual participants' criteria, as reflected in their probability of responding "bot" across 60 personas. As in Studies 1 and 2, participants in the control condition exhibited a greater tendency to label stimuli as humans, affirming H4. As in Study 2, the aid group's criterion indicated less willingness to accept a persona as a human, affirming H5. However, the training group's criterion was no different from that of the control group, contrary to H6. These results are reflected in the simple group effects in Model 1.

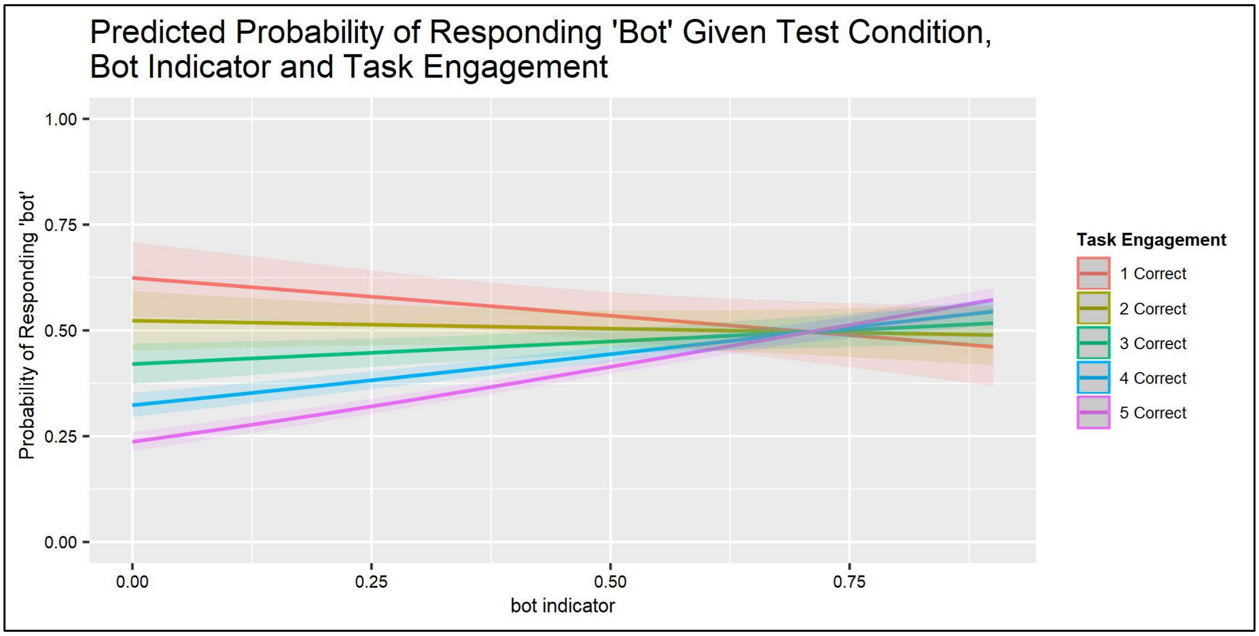
Figure 2



Note. An unbiased criterion would equal a 50% probability of responding 'bot' or 'human.' Criteria below 50% indicate a tendency to respond 'human.' Criteria above 50% indicate a tendency to respond 'bot.'

Controls. In Model 1, both controls were significant predictors of sensitivity (as seen in the interactions, BIxTE and BIxSRO) and criterion (as seen in the main effects for TE and SRO). As Stimulus Presentation Order increased, so did sensitivity, contrary to the predicted fatigue effect (H7), while the tendency to respond "human" decreased (contrary to H9). Participants with higher Task Engagement scores demonstrate greater sensitivity (affirming H8) and less tendency to respond "human" (contrary to H9). Figure 3 shows these effects. The effect sizes were small.

Figure 3



Note. Task engagement was measured as the total number of correct attention checks (one prior, three during the experimental procedure, and one post).

Social Media Experience

Model 2 adds self-reported Social Media Experience (SME) to the regression. As in Studies 2 and 3, it has little value as a predictor, affirming H10. It is unrelated to the criterion either overall (as seen in the nonsignificant main effect) or by condition (as seen in the nonsignificant interactions with the Training Group and Aid Group). It is unrelated to sensitivity overall (as seen in the lack of interaction with BI) and has a small, uninterpretable three-way interaction with BI and Test Condition (p =0.04). Its addition does not change the patterns in Model 1.

Analytical Reasoning Ability (CRT)

. Model 3 adds scores on the Cognitive Reflection Test (CRT). They were unrelated to the criterion (as seen in the nonsignificant CRT main effect) or sensitivity (as seen in the nonsignificant CRTxBI interaction). That result is contrary to H11a and H11b, which reflect prior research, but consistent with our other results. Exploratory analyses found a significant interaction with the test condition. Participants with higher CRT scores were more likely to identify a stimulus as a bot in the training group than in the aid and control groups, suggesting that they might have been better able to take advantage of the training. These effects vary once political identity and differences are included in Models 4 and 5 (discussed below).

Political Values and Political Differences

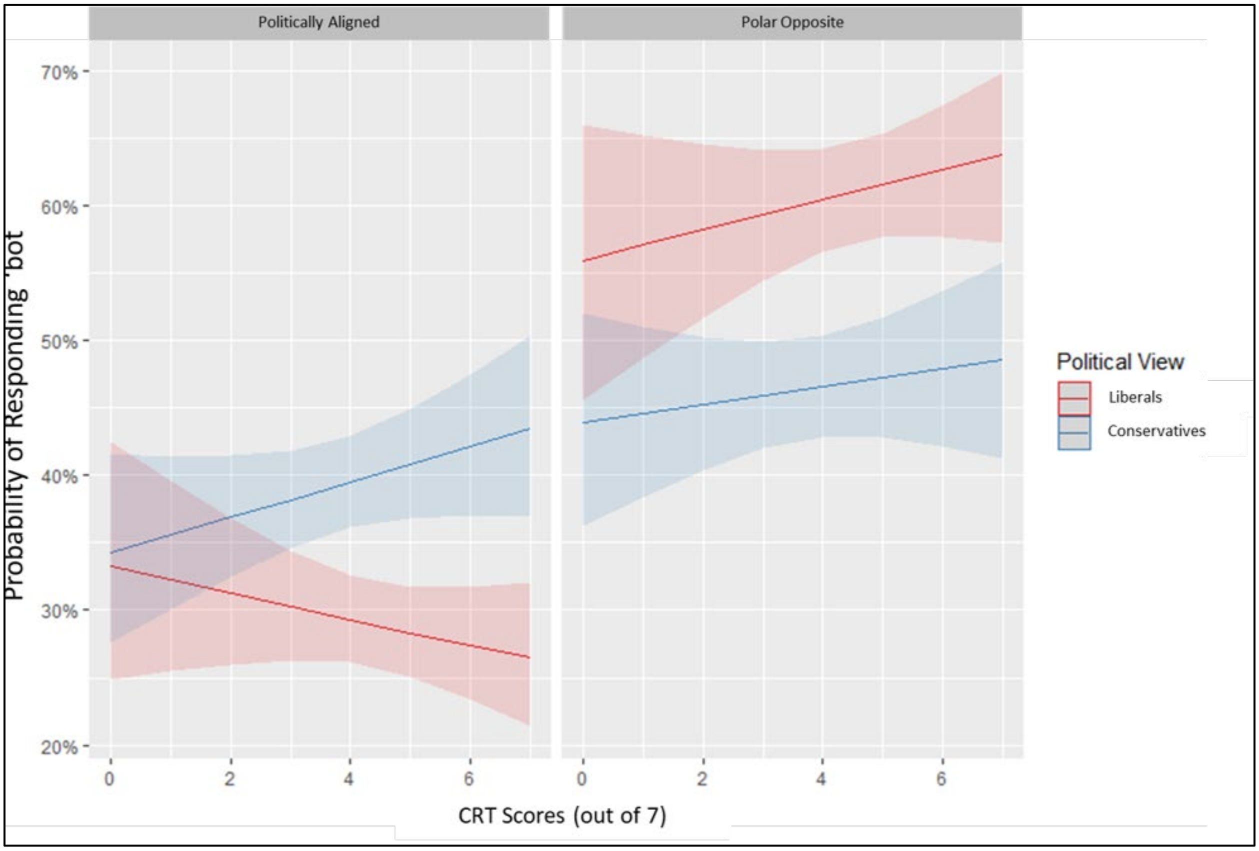
Criterion. Model 4 shows that participants' political values (PV) were unrelated to their criterion for saying "bot" or their sensitivity to whether stimuli were bots, as reflected in the non-significant main effect and interaction with BI. We had no prediction here but included Model 4 to examine the effects of PV per se before considering political differences (PD).

Model 5 adds the political difference (PD) between a participant's political identity and that of a persona. The significant main effect ($p < 0.001$) of PD reflects the myside bias seen in Studies 1 and 2, affirming H12. Participants have a lower threshold for treating personas as a bot, the greater the difference between their political views.

We found a relationship between CRT scores and myside bias that supports H13. Exploratory analyses found that the size of myside bias was related to a complex interaction between political views, political difference, and analytical reasoning (PDxPVxCRT; $p = 0.029$), consistent with the non-significant PVxPD interactions in previous studies. Liberal participants' thresholds for responding 'bot' when considering conservatives decreased to a greater extent than conservatives' threshold when considering liberals. The nonsignificant BI x PV x PD x CRT interaction ($p = 0.11$) reflected a weak but intriguing tendency for liberals to be more sensitive when viewing liberals and conservatives to be less sensitive when viewing conservatives.

Figure 4

Probability of Responding 'Bot' Given Political Views, Political Differences, and Analytical Reasoning Ability



Behavioral Responses

Study 3 included two behavioral responses, participants' willingness to retweet a message if they agreed with its content and willingness to pay for a service like that offered in the aid condition. We analyze each for insight into the potential behavioral consequences of participants' judgments and the impact of the interventions on them.

Willingness to retweet

The dependent variable for this analysis was the probability of participants responding 'yes' when asked if they would retweet content from a persona if they agreed with its content. On average, participants would retweet only 17% of all personas. The percentage varied across conditions, with aid group participants retweeting 14% of personas, control group participants retweeting 15%, and training group participants retweeting 22% ($p < 0.001$).

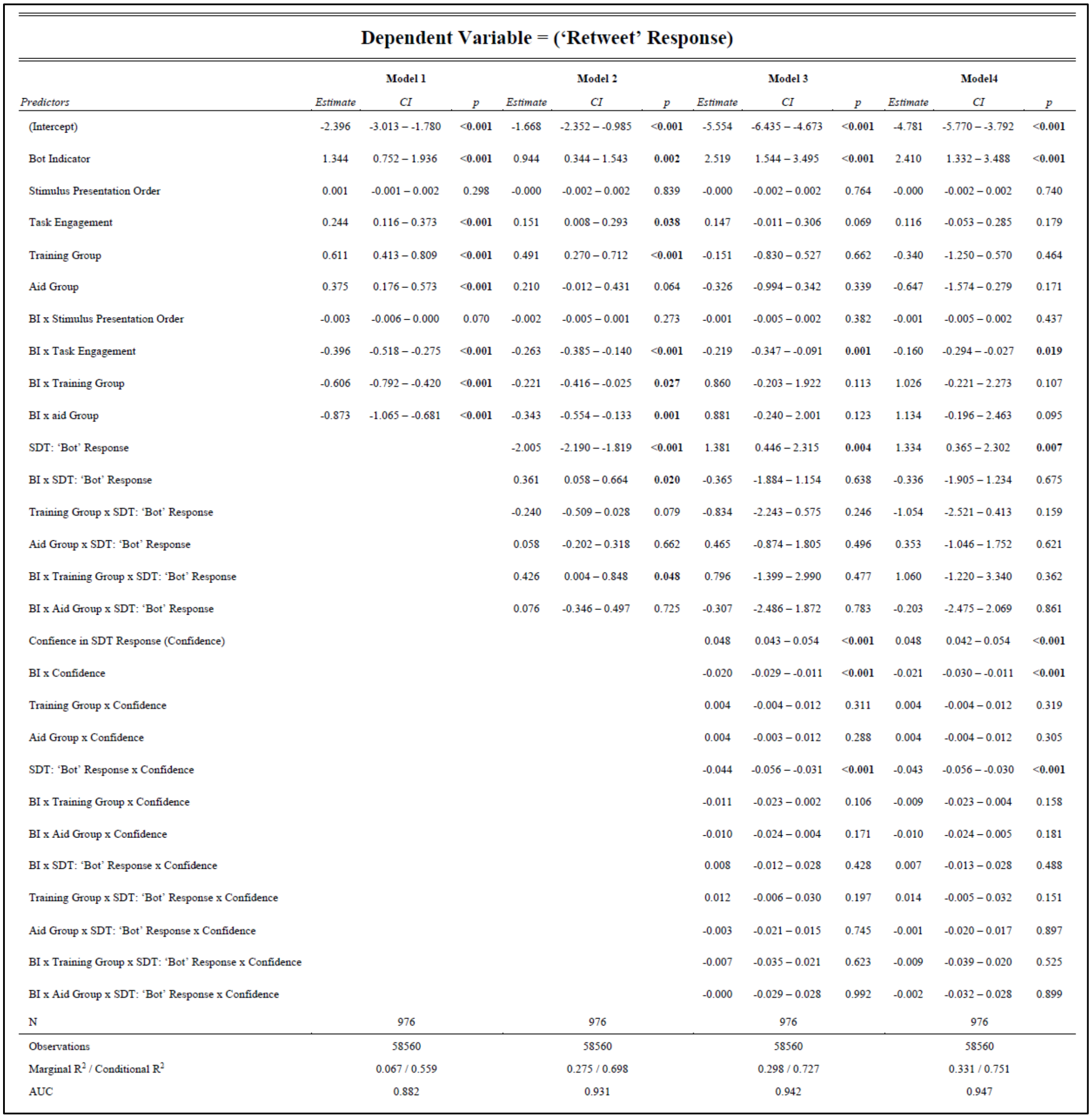
Our pre-planned analysis employed general linear mixed-effects regression to predict that probability based on the fixed effects of the experimental condition (control, aid, training), stimuli attributes (bot indicator score, political tone), individual attributes (social media experience, CRT, political view, political difference), response (bot or human), and confidence.

All models employed a Probit link function. To predict a probability of responding 'yes,' z-scores are converted to probabilities by taking the inverse of Phi. Random effects for each model coefficient were estimated using a multivariate Gaussian distribution using the arm package in R. Tables 2a and 2b present these results, all of which are exploratory analyses.

Model 1 uses bot indicator (BI), the controls (SPO, TE), and test groups as predictors; Model 2 adds the participants' bot/human response; Model 3 adds participants' confidence in that response; Model 4 includes the individual attributes found (above) to affect sensitivity or criterion, separately or in interactions: CRT, political view (PV), and political differences (PD).

Table 2a

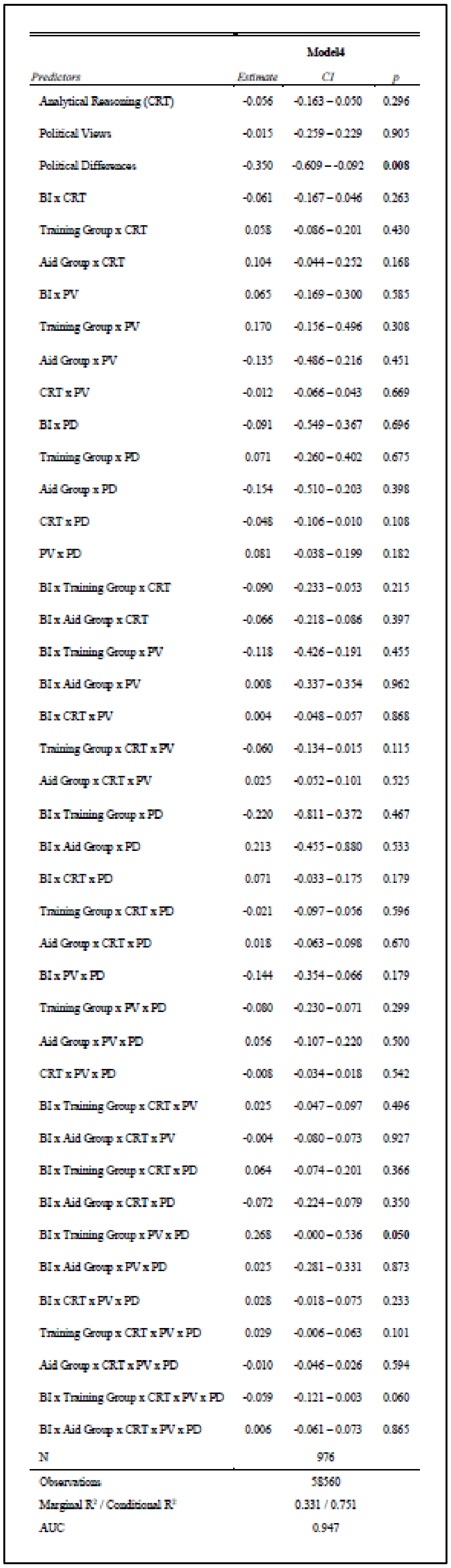
General Linear Mixed Effects Probit Regression Models, Predicting the Probability of Retweeting Content from a Persona



Dependent Variable = ('Retweet' Response)

	Model 1			**Model 2**			**Model 3**			**Model4**		
Predictors	*Estimate*	*CI*	*p*	*Estimate*	*CI*	*p*	*Estimate*	*CI*	*p*	*Estimate*	*CI*	*p*
(Intercept)	-2.396	-3.013 – -1.780	**<0.001**	-1.668	-2.352 – -0.985	**<0.001**	-5.554	-6.435 – -4.673	**<0.001**	-4.781	-5.770 – -3.792	**<0.001**
Bot Indicator	1.344	0.752 – 1.936	**<0.001**	0.944	0.344 – 1.543	**0.002**	2.519	1.544 – 3.495	**<0.001**	2.410	1.332 – 3.488	**<0.001**
Stimulus Presentation Order	0.001	-0.001 – 0.002	0.298	-0.000	-0.002 – 0.002	0.839	-0.000	-0.002 – 0.002	0.764	-0.000	-0.002 – 0.002	0.740
Task Engagement	0.244	0.116 – 0.373	**<0.001**	0.151	0.008 – 0.293	**0.038**	0.147	-0.011 – 0.306	0.069	0.116	-0.053 – 0.285	0.179
Training Group	0.611	0.413 – 0.809	**<0.001**	0.491	0.270 – 0.712	**<0.001**	-0.151	-0.830 – 0.527	0.662	-0.340	-1.250 – 0.570	0.464
Aid Group	0.375	0.176 – 0.573	**<0.001**	0.210	-0.012 – 0.431	0.064	-0.326	-0.994 – 0.342	0.339	-0.647	-1.574 – 0.279	0.171
BI x Stimulus Presentation Order	-0.003	-0.006 – 0.000	0.070	-0.002	-0.005 – 0.001	0.273	-0.001	-0.005 – 0.002	0.382	-0.001	-0.005 – 0.002	0.437
BI x Task Engagement	-0.396	-0.518 – -0.275	**<0.001**	-0.263	-0.385 – -0.140	**<0.001**	-0.219	-0.347 – -0.091	**0.001**	-0.160	-0.294 – -0.027	**0.019**
BI x Training Group	-0.606	-0.792 – -0.420	**<0.001**	-0.221	-0.416 – -0.025	**0.027**	0.860	-0.203 – 1.922	0.113	1.026	-0.221 – 2.273	0.107
BI x aid Group	-0.873	-1.065 – -0.681	**<0.001**	-0.343	-0.554 – -0.133	**0.001**	0.881	-0.240 – 2.001	0.123	1.134	-0.196 – 2.463	0.095
SDT: 'Bot' Response				-2.005	-2.190 – -1.819	**<0.001**	1.381	0.446 – 2.315	**0.004**	1.334	0.365 – 2.302	**0.007**
BI x SDT: 'Bot' Response				0.361	0.058 – 0.664	**0.020**	-0.365	-1.884 – 1.154	0.638	-0.336	-1.905 – 1.234	0.675
Training Group x SDT: 'Bot' Response				-0.240	-0.509 – 0.028	0.079	-0.834	-2.243 – 0.575	0.246	-1.054	-2.521 – 0.413	0.159
Aid Group x SDT: 'Bot' Response				0.058	-0.202 – 0.318	0.662	0.465	-0.874 – 1.805	0.496	0.353	-1.046 – 1.752	0.621
BI x Training Group x SDT: 'Bot' Response				0.426	0.004 – 0.848	**0.048**	0.796	-1.399 – 2.990	0.477	1.060	-1.220 – 3.340	0.362
BI x Aid Group x SDT: 'Bot' Response				0.076	-0.346 – 0.497	0.725	-0.307	-2.486 – 1.872	0.783	-0.203	-2.475 – 2.069	0.861
Confience in SDT Response (Confidence)							0.048	0.043 – 0.054	**<0.001**	0.048	0.042 – 0.054	**<0.001**
BI x Confidence							-0.020	-0.029 – -0.011	**<0.001**	-0.021	-0.030 – -0.011	**<0.001**
Training Group x Confidence							0.004	-0.004 – 0.012	0.311	0.004	-0.004 – 0.012	0.319
Aid Group x Confidence							0.004	-0.003 – 0.012	0.288	0.004	-0.004 – 0.012	0.305
SDT: 'Bot' Response x Confidence							-0.044	-0.056 – -0.031	**<0.001**	-0.043	-0.056 – -0.030	**<0.001**
BI x Training Group x Confidence							-0.011	-0.023 – 0.002	0.106	-0.009	-0.023 – 0.004	0.158
BI x Aid Group x Confidence							-0.010	-0.024 – 0.004	0.171	-0.010	-0.024 – 0.005	0.181
BI x SDT: 'Bot' Response x Confidence							0.008	-0.012 – 0.028	0.428	0.007	-0.013 – 0.028	0.488
Training Group x SDT: 'Bot' Response x Confidence							0.012	-0.006 – 0.030	0.197	0.014	-0.005 – 0.032	0.151
Aid Group x SDT: 'Bot' Response x Confidence							-0.003	-0.021 – 0.015	0.745	-0.001	-0.020 – 0.017	0.897
BI x Training Group x SDT: 'Bot' Response x Confidence							-0.007	-0.035 – 0.021	0.623	-0.009	-0.039 – 0.020	0.525
BI x Aid Group x SDT: 'Bot' Response x Confidence							-0.000	-0.029 – 0.028	0.992	-0.002	-0.032 – 0.028	0.899
N	976			976			976			976		
Observations	58560			58560			58560			58560		
Marginal R^2 / Conditional R^2	0.067 / 0.559			0.275 / 0.698			0.298 / 0.727			0.331 / 0.751		
AUC	0.882			0.931			0.942			0.947		

Table 2b

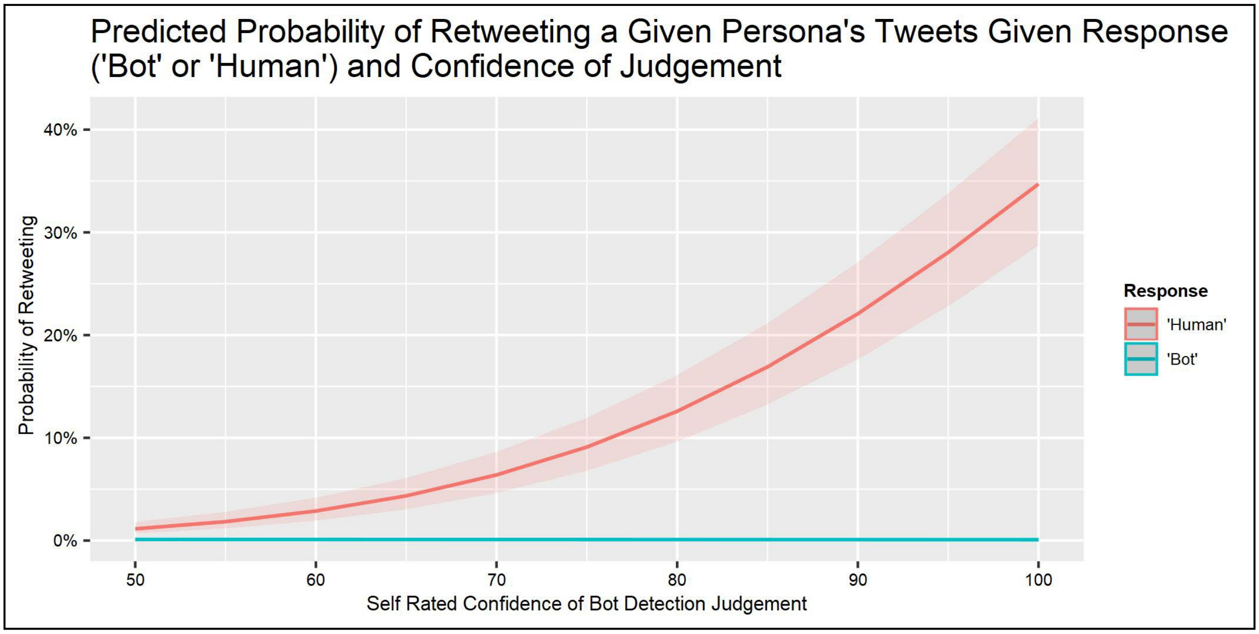


	Model4		
Predictors	*Estimate*	*CI*	*p*
Analytical Reasoning (CRT)	-0.056	-0.163 – 0.050	0.296
Political Views	-0.015	-0.259 – 0.229	0.905
Political Differences	-0.350	-0.609 – -0.092	**0.008**
BI x CRT	-0.061	-0.167 – 0.046	0.263
Training Group x CRT	0.058	-0.086 – 0.201	0.430
Aid Group x CRT	0.104	-0.044 – 0.252	0.168
BI x PV	0.065	-0.169 – 0.300	0.585
Training Group x PV	0.170	-0.156 – 0.496	0.308
Aid Group x PV	-0.135	-0.486 – 0.216	0.451
CRT x PV	-0.012	-0.066 – 0.043	0.669
BI x PD	-0.091	-0.549 – 0.367	0.696
Training Group x PD	0.071	-0.260 – 0.402	0.675
Aid Group x PD	-0.154	-0.510 – 0.203	0.398
CRT x PD	-0.048	-0.106 – 0.010	0.108
PV x PD	0.081	-0.038 – 0.199	0.182
BI x Training Group x CRT	-0.090	-0.233 – 0.053	0.215
BI x Aid Group x CRT	-0.066	-0.218 – 0.086	0.397
BI x Training Group x PV	-0.118	-0.426 – 0.191	0.455
BI x Aid Group x PV	0.008	-0.337 – 0.354	0.962
BI x CRT x PV	0.004	-0.048 – 0.057	0.868
Training Group x CRT x PV	-0.060	-0.134 – 0.015	0.115
Aid Group x CRT x PV	0.025	-0.052 – 0.101	0.525
BI x Training Group x PD	-0.220	-0.811 – 0.372	0.467
BI x Aid Group x PD	0.213	-0.455 – 0.880	0.533
BI x CRT x PD	0.071	-0.033 – 0.175	0.179
Training Group x CRT x PD	-0.021	-0.097 – 0.056	0.596
Aid Group x CRT x PD	0.018	-0.063 – 0.098	0.670
BI x PV x PD	-0.144	-0.354 – 0.066	0.179
Training Group x PV x PD	-0.080	-0.230 – 0.071	0.299
Aid Group x PV x PD	0.056	-0.107 – 0.220	0.500
CRT x PV x PD	-0.008	-0.034 – 0.018	0.542
BI x Training Group x CRT x PV	0.025	-0.047 – 0.097	0.496
BI x Aid Group x CRT x PV	-0.004	-0.080 – 0.073	0.927
BI x Training Group x CRT x PD	0.064	-0.074 – 0.201	0.366
BI x Aid Group x CRT x PD	-0.072	-0.224 – 0.079	0.350
BI x Training Group x PV x PD	0.268	-0.000 – 0.536	**0.050**
BI x Aid Group x PV x PD	0.025	-0.281 – 0.331	0.873
BI x CRT x PV x PD	0.028	-0.018 – 0.075	0.233
Training Group x CRT x PV x PD	0.029	-0.006 – 0.063	0.101
Aid Group x CRT x PV x PD	-0.010	-0.046 – 0.026	0.594
BI x Training Group x CRT x PV x PD	-0.059	-0.121 – 0.003	0.060
BI x Aid Group x CRT x PV x PD	0.006	-0.061 – 0.073	0.865
N		976	
Observations		58560	
Marginal R^2 / Conditional R^2		0.331 / 0.751	
AUC		0.947	

The intercept remained negative across the four models (p < 0.001), indicating a very high criterion for retweeting, regardless of other factors.

Figure 5, which pools across test conditions, given the similar retweet rates, reveals that when participants respond 'human,' their probability of retweeting content increases dramatically with their confidence in that judgment (p = 0.007). When they respond “bot,” they never retweet, even if they are not confident at all. The interaction of bot/human response with confidence reveals that if a participant judged a persona as 'human,' their probability of retweeting their content increased as the confidence in their judgment increased. (p < 0.001).

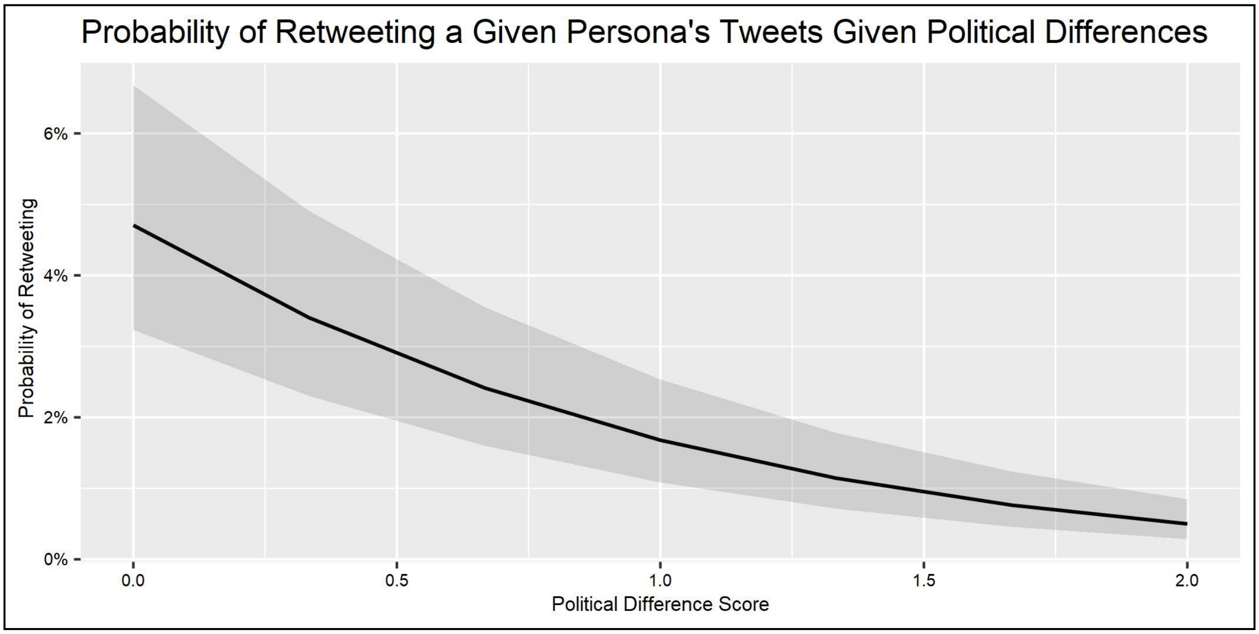
Figure 5



Overall, participants expressed moderate confidence in their bot/human judgments. Pooling groups and stimuli: M=78.7% (SD=15.1%), on the 50-100% scale. Given with the comparable 65.1% accuracy rate, participants were overconfident overall (78.7%-65.1% = 13.6%). Mean confidence varied some (p<???) by condition (control = 76.2%, training = 81.9%, aid = 77.9%). Mean accuracy varied much more (control = 56.4%, training = 67.9%, aid = 70.6%). As a result, overconfidence was much greater in the control condition (19.8%) than the aid condition (7.3%), with the training condition in between (14%).

Figure 6 shows myside bias, as reflected in decreasing willingness to retweet content as the difference in political views grows (p < 0.001).

Figure 6



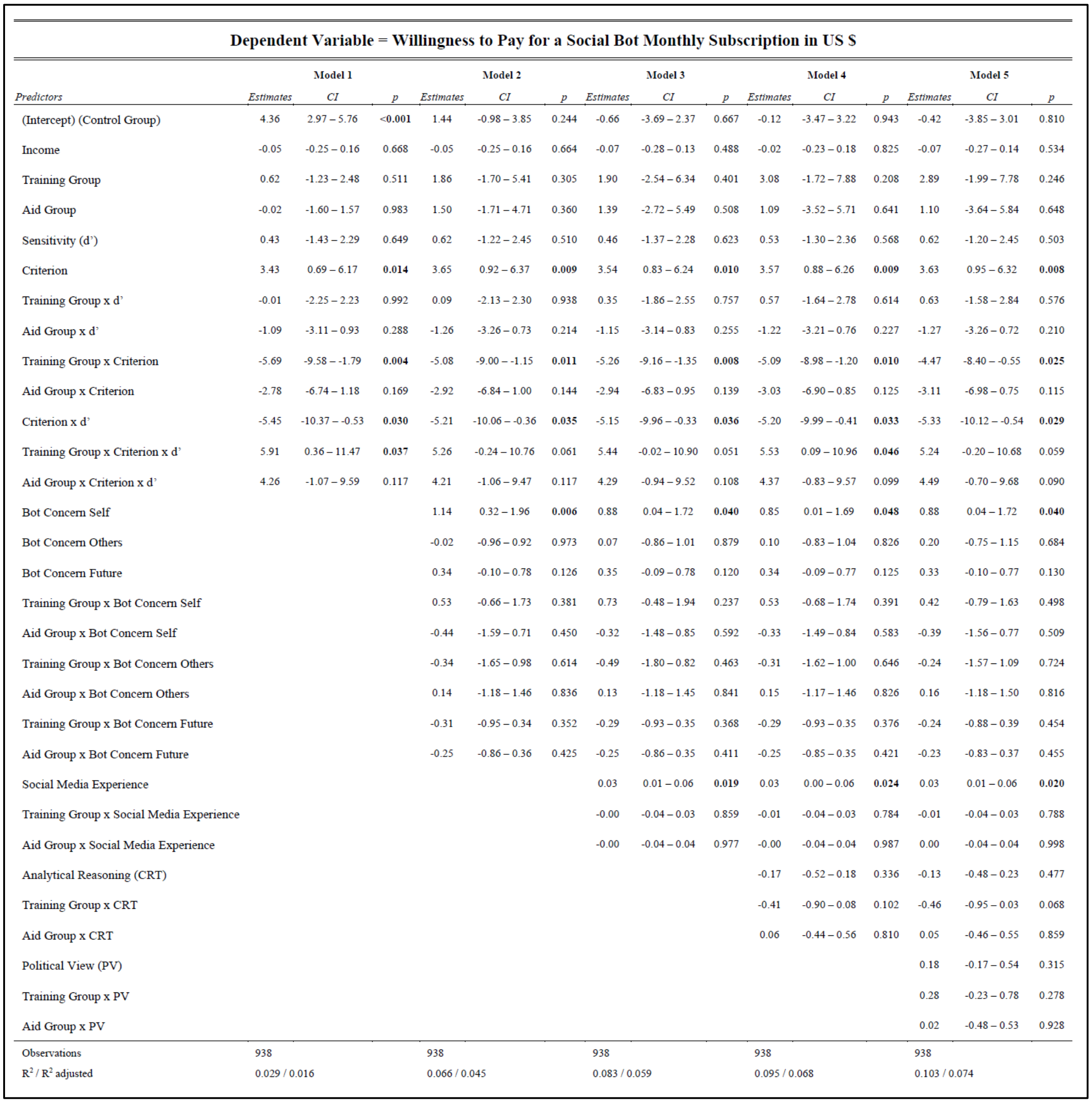
Willingness to Pay - Monetary

After completing the primary tasks, participants were asked, “What is the maximum amount you would be willing to pay monthly for a social bot detection tool that told you which online personas are social bots?” They could enter any amount with no upper bounds, rounded to $1 intervals. Nine participant responses were eliminated because they surpassed 99% of response values (i.e., $100,000 vs. an average of $2).

Our pre-planned analysis employed the same linear regression approach as in study 2 to consider the predictive value of the fixed effects of the experimental condition (control, aid, training), individual attributes (social media experience, cognitive reflection ability, bot concerns, and social bot detection performance estimates observed in the study (i.e., d', and criterion) in determining participants' self-reported WTP. Participants, as part of the demographic survey, reported their annual income. Annual incomes were normally distributed, ranging from "less than 10K" to "over 150K," with the median between 50K-75K. These values were binned linearly and used as a covariate. Across models, it was not a significant predictor. Thus, willingness to pay was not related to the ability to pay. Table 3 presents the results.

Table 3

Linear Models Predicting Willingness to Pay for a Monthly Social Bot Detection Service.



Dependent Variable = Willingness to Pay for a Social Bot Monthly Subscription in US $

	Model 1			Model 2			Model 3			Model 4			Model 5		
Predictors	*Estimates*	*CI*	*p*	*Estimates*	*CI*	*p*	*Estimates*	*CI*	*p*	*Estimates*	*CI*	*p*	*Estimates*	*CI*	*p*
(Intercept) (Control Group)	4.36	2.97 – 5.76	**<0.001**	1.44	-0.98 – 3.85	0.244	-0.66	-3.69 – 2.37	0.667	-0.12	-3.47 – 3.22	0.943	-0.42	-3.85 – 3.01	0.810
Income	-0.05	-0.25 – 0.16	0.668	-0.05	-0.25 – 0.16	0.664	-0.07	-0.28 – 0.13	0.488	-0.02	-0.23 – 0.18	0.825	-0.07	-0.27 – 0.14	0.534
Training Group	0.62	-1.23 – 2.48	0.511	1.86	-1.70 – 5.41	0.305	1.90	-2.54 – 6.34	0.401	3.08	-1.72 – 7.88	0.208	2.89	-1.99 – 7.78	0.246
Aid Group	-0.02	-1.60 – 1.57	0.983	1.50	-1.71 – 4.71	0.360	1.39	-2.72 – 5.49	0.508	1.09	-3.52 – 5.71	0.641	1.10	-3.64 – 5.84	0.648
Sensitivity (d’)	0.43	-1.43 – 2.29	0.649	0.62	-1.22 – 2.45	0.510	0.46	-1.37 – 2.28	0.623	0.53	-1.30 – 2.36	0.568	0.62	-1.20 – 2.45	0.503
Criterion	3.43	0.69 – 6.17	**0.014**	3.65	0.92 – 6.37	**0.009**	3.54	0.83 – 6.24	**0.010**	3.57	0.88 – 6.26	**0.009**	3.63	0.95 – 6.32	**0.008**
Training Group x d’	-0.01	-2.25 – 2.23	0.992	0.09	-2.13 – 2.30	0.938	0.35	-1.86 – 2.55	0.757	0.57	-1.64 – 2.78	0.614	0.63	-1.58 – 2.84	0.576
Aid Group x d’	-1.09	-3.11 – 0.93	0.288	-1.26	-3.26 – 0.73	0.214	-1.15	-3.14 – 0.83	0.255	-1.22	-3.21 – 0.76	0.227	-1.27	-3.26 – 0.72	0.210
Training Group x Criterion	-5.69	-9.58 – -1.79	**0.004**	-5.08	-9.00 – -1.15	**0.011**	-5.26	-9.16 – -1.35	**0.008**	-5.09	-8.98 – -1.20	**0.010**	-4.47	-8.40 – -0.55	**0.025**
Aid Group x Criterion	-2.78	-6.74 – 1.18	0.169	-2.92	-6.84 – 1.00	0.144	-2.94	-6.83 – 0.95	0.139	-3.03	-6.90 – 0.85	0.125	-3.11	-6.98 – 0.75	0.115
Criterion x d’	-5.45	-10.37 – -0.53	**0.030**	-5.21	-10.06 – -0.36	**0.035**	-5.15	-9.96 – -0.33	**0.036**	-5.20	-9.99 – -0.41	**0.033**	-5.33	-10.12 – -0.54	**0.029**
Training Group x Criterion x d’	5.91	0.36 – 11.47	**0.037**	5.26	-0.24 – 10.76	0.061	5.44	-0.02 – 10.90	0.051	5.53	0.09 – 10.96	**0.046**	5.24	-0.20 – 10.68	0.059
Aid Group x Criterion x d’	4.26	-1.07 – 9.59	0.117	4.21	-1.06 – 9.47	0.117	4.29	-0.94 – 9.52	0.108	4.37	-0.83 – 9.57	0.099	4.49	-0.70 – 9.68	0.090
Bot Concern Self				1.14	0.32 – 1.96	**0.006**	0.88	0.04 – 1.72	**0.040**	0.85	0.01 – 1.69	**0.048**	0.88	0.04 – 1.72	**0.040**
Bot Concern Others				-0.02	-0.96 – 0.92	0.973	0.07	-0.86 – 1.01	0.879	0.10	-0.83 – 1.04	0.826	0.20	-0.75 – 1.15	0.684
Bot Concern Future				0.34	-0.10 – 0.78	0.126	0.35	-0.09 – 0.78	0.120	0.34	-0.09 – 0.77	0.125	0.33	-0.10 – 0.77	0.130
Training Group x Bot Concern Self				0.53	-0.66 – 1.73	0.381	0.73	-0.48 – 1.94	0.237	0.53	-0.68 – 1.74	0.391	0.42	-0.79 – 1.63	0.498
Aid Group x Bot Concern Self				-0.44	-1.59 – 0.71	0.450	-0.32	-1.48 – 0.85	0.592	-0.33	-1.49 – 0.84	0.583	-0.39	-1.56 – 0.77	0.509
Training Group x Bot Concern Others				-0.34	-1.65 – 0.98	0.614	-0.49	-1.80 – 0.82	0.463	-0.31	-1.62 – 1.00	0.646	-0.24	-1.57 – 1.09	0.724
Aid Group x Bot Concern Others				0.14	-1.18 – 1.46	0.836	0.13	-1.18 – 1.45	0.841	0.15	-1.17 – 1.46	0.826	0.16	-1.18 – 1.50	0.816
Training Group x Bot Concern Future				-0.31	-0.95 – 0.34	0.352	-0.29	-0.93 – 0.35	0.368	-0.29	-0.93 – 0.35	0.376	-0.24	-0.88 – 0.39	0.454
Aid Group x Bot Concern Future				-0.25	-0.86 – 0.36	0.425	-0.25	-0.86 – 0.35	0.411	-0.25	-0.85 – 0.35	0.421	-0.23	-0.83 – 0.37	0.455
Social Media Experience							0.03	0.01 – 0.06	**0.019**	0.03	0.00 – 0.06	**0.024**	0.03	0.01 – 0.06	**0.020**
Training Group x Social Media Experience							-0.00	-0.04 – 0.03	0.859	-0.01	-0.04 – 0.03	0.784	-0.01	-0.04 – 0.03	0.788
Aid Group x Social Media Experience							-0.00	-0.04 – 0.04	0.977	-0.00	-0.04 – 0.04	0.987	0.00	-0.04 – 0.04	0.998
Analytical Reasoning (CRT)										-0.17	-0.52 – 0.18	0.336	-0.13	-0.48 – 0.23	0.477
Training Group x CRT										-0.41	-0.90 – 0.08	0.102	-0.46	-0.95 – 0.03	0.068
Aid Group x CRT										0.06	-0.44 – 0.56	0.810	0.05	-0.46 – 0.55	0.859
Political View (PV)													0.18	-0.17 – 0.54	0.315
Training Group x PV													0.28	-0.23 – 0.78	0.278
Aid Group x PV													0.02	-0.48 – 0.53	0.928
Observations	938			938			938			938			938		
R^2 / R^2 adjusted	0.029 / 0.016			0.066 / 0.045			0.083 / 0.059			0.095 / 0.068			0.103 / 0.074		

Participants' willingness to pay for a social bot detection service did not vary by condition (Model 1). Thus, exposure to aid did not affect willingness to pay for a related service. The mean WTP across all participants was $4.32. Figure 7 shows WTP did vary with sensitivity and criterion. Pooling test conditions, participants with lower thresholds for responding bot were willing to pay more than participants with higher thresholds for responding bot.

Figure 7

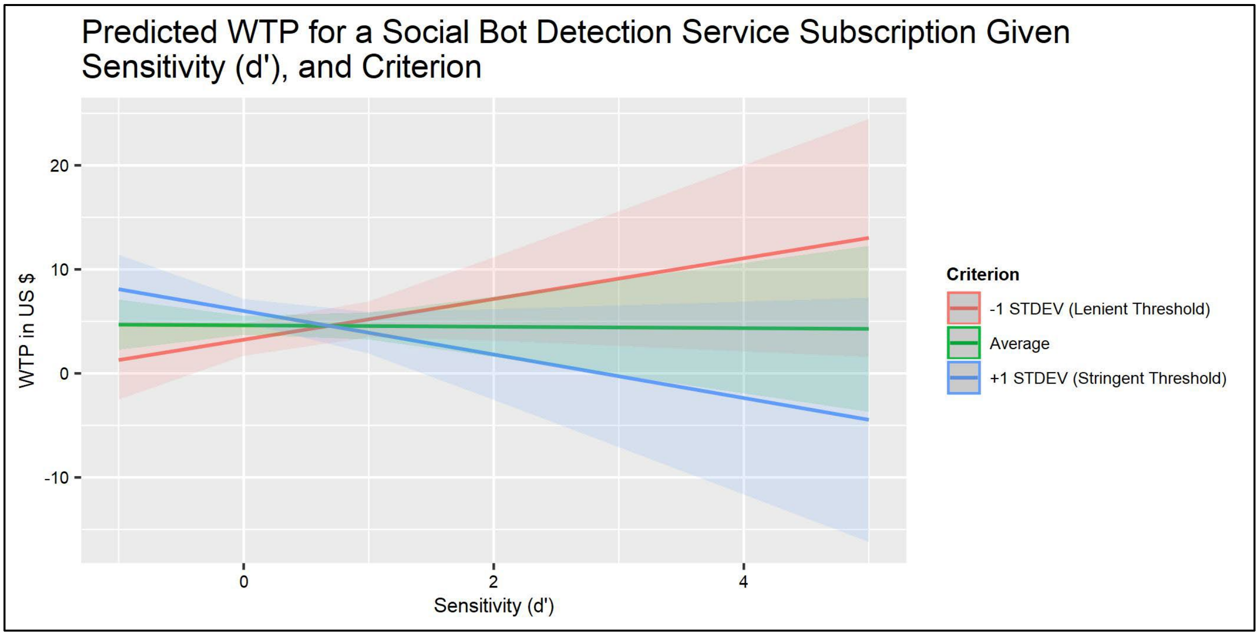
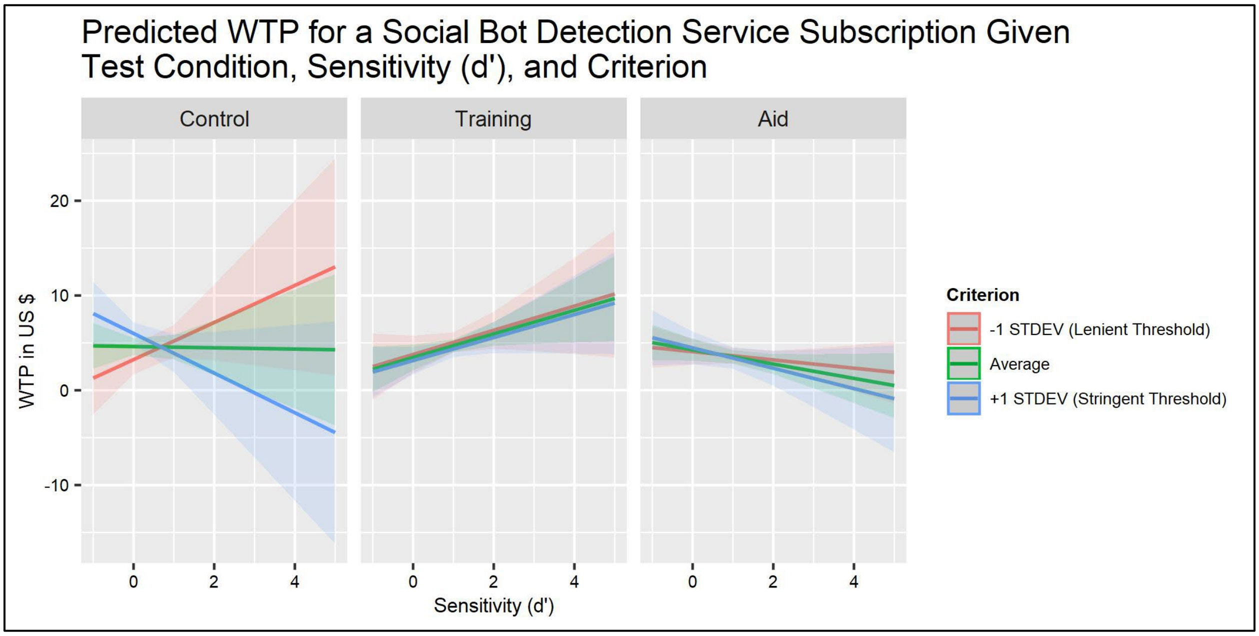


Figure 8 shows the three-way interaction of test condition, sensitivity, and the leniency of one's criterion. In the aid group, participants with higher sensitivity were less willing to pay for aid. In the training group, they were willing to pay more. In the control condition, participants with lower thresholds for responding bot and higher sensitivity were willing to pay more than participants with higher thresholds for responding bot.

Figure 8

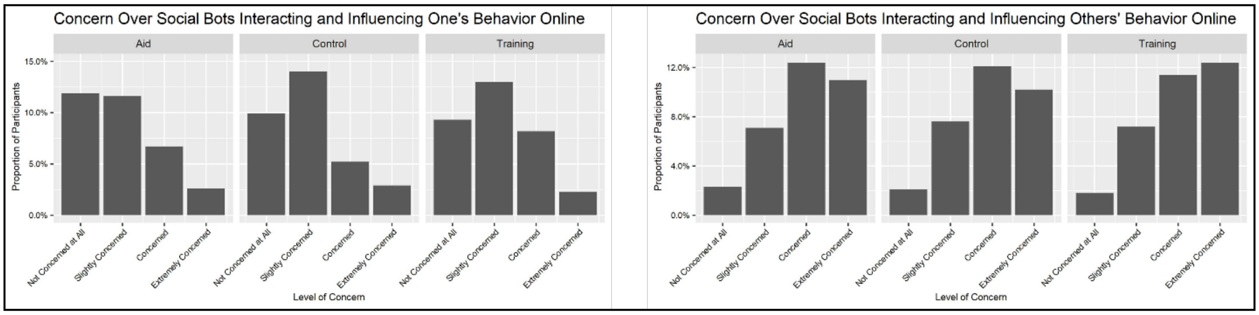


The most significant predictor among the individual attributes was social media experience (p=0.02). As in Study 2, participants with more self-reported experience were willing to pay more (p = 0.02).

Social Bot Concerns

Figure 9 shows participants' responses to "How concerned are you about social bots affecting your [others'] behavior online?" As in Study 2, participants were more concerned about others being influenced by social bots than being affected themselves (1.99 vs. 1.04, on a 0-3 scale; p < 0.001). Those with the most concern were willing to pay more for a social bot detection service (p = 0.04).

Figure 9



Note. We asked participants, "How concerned are you about social bots affecting YOUR/OTHERS' behavior online?"

Discussion

The control group in Study 3 replicated the patterns from Studies 1 and 2. Under these experimental conditions, people are somewhat sensitive to whether personas are social bots and have a decision rule that assumes that personas are humans rather than bots. They are also subject to myside bias, such that their criterion leans toward assuming the personas are humans when they share political views and bots when they do not. Political views per se were, again, unrelated to sensitivity or criterion. However, political views interact with other variables in complex ways. Although not central to the present study, these interactions, revealed by this novel task, bear further attention for understanding the dynamics of online behavior.

The aid group replicated the Study 2 finding that people will and can use this algorithmic aid, contrary to the reluctance observed in many other settings. Providing a bot indicator score increased their sensitivity and made them more cautious about treating personas as humans. It did not reduce their myside bias. The failure of the reminder condition in Study 2 indicated that the aid's content made the difference, not just its presence.

The training video condition was designed for situations where algorithmic interventions were unavailable. It described and demonstrated two cues with known predictive value and potentially enough intuitive resonance to become part of Twitter users' natural heuristics. The training video improved sensitivity to a degree approaching that of the aid. It did not affect participants' criteria, despite warning them about the motives of social bot creators.

Again, participants' behavioral responses were generally consistent with their judgments of the personas. Participants across conditions had a low probability of retweeting any given persona (approximately 15% or less). That probability plummeted if they labeled a persona a bot. People with the most experience, and likely daily engagement on social media, were more willing to pay for aid in detecting bots. So too were those with the most concern about social bots influencing their online behaviors. The interaction of intervention with sensitivity and one's response threshold deserves further research.

The individual difference variables made weak contributions to predicting performance, with potential implications for theory and practice. Cognitive Reflection Test (CRT) scores were primarily unrelated to sensitivity, criteria, or behavior, except for their interaction with political views and political differences. We suspect that CRT does not capture the complex, substantively informed inferences that bot detection requires despite the plausible assumption that people who score higher should perform better. Self-reported social media experience continued to have weak, inconsistent relationships with other measures. We suspect that our measure did not specifically address Twitter use. We also suspect that experience supports performance by providing substantive knowledge and undermines it by conferring a false sense of mastery. Political differences, but not political views, were a strong predictor of decision criteria.

Implications and Future Work

Our findings provide evidence for two strategies for improving social bot detection: a bot indicator score aid and a brief training video. The aid affected users' sensitivity and criterion in appropriate ways (assuming that control participants were too willing to accept personas as humans). The video training increased sensitivity without affecting the criterion, despite emphasizing the manipulative strategies of social bot developers. In all conditions, participants who responded 'bot' were far less likely to retweet content from a persona, even if they agreed with it. Thus, any intervention that improves social bot detection ability through training should reduce the impact of social bots.

In most conditions, performance increased with task engagement and decreased with stimulus presentation order (and, presumably, fatigue). As a result, the improved performance observed here, with the aid and training, might not be sustained over longer Twitter sessions.

Avoiding long sessions would be good advice. We cannot know, without retesting, how long the effects of the video training would last or whether continuing exposure to the aid would increase its effectiveness or lead to habituation. A reason for optimism might be found in the sustained success of a single training session for phishing email detection (Sarno, McPherson & Neider, 2022).

Another task for future research is developing interventions that reduce myside bias, a domain where other interventions, too, have had mixed results (Stanovich, 2021; Drummond, & Fischhoff, 2019; Jurkovič, 2016Pennycook, Epstein, Mosleh, Arechar, Eckles, & Rand, 2021; Pennycook, McPhetres, Zhang, Lu, & Rand, 2020). Those interventions have typically focused on increasing sensitivity by improving detection ability, as with our aid and training. Our video noted the motives of social bot detectors. It might have been more effective had it said more about the consequences of falling prey to bots.

The training video is, we believe, the first of its kind in this domain. Its success suggests pursuing the strategy further. First, identify statistically valid cues, then help users incorporate them into their intuitive heuristics by explaining their rationale. Because social bot developers seek to amplify their narrative, excessive tweeting suggests a bot. Because social bot developers seek to share information with an extensive social network, a high follow-to-follower ratio suggests a bot. A better understanding of users' mental models might improve the use of these cues. Additional statistical analysis might identify better cues.

Heuristic training can also empower users when algorithms are unavailable, including when bots are too new to be characterized statistically. As adversarial social bot developers seek to circumvent new approaches to detection, these predictors and heuristics may need regular updating. Indeed, some social bot developers have reduced their rate to avoid detection (Orabi, Mouheb, Al Aghbari, & Kamel, 2020; Cresci, 2019; Cresci, 2020). In this game of cat and mouse between social bot developers and those seeking to detect them, there may be a need for continuing user-focused development of interventions to help social media users protect themselves. That development will require collaboration between computer scientists, behavioral decision researchers, policymakers, and system managers, to understand the evolving threat landscape, devise responses, and assess the remaining vulnerabilities.

Key Points

- Twitter users have some ability to detect social bots while tending to treat uncertain personas as humans rather than bots.
- Providing an AI-produced bot indicator score improves users' ability to detect and makes them more cautious about accepting personas as humans.
- A short training video improves users' ability to detect bots without affecting their decisions about uncertain stimuli.
- Users rarely report willingness to retweet tweets, especially as their suspicions of a persona being a bot increase.
- Users are subject to myside bias, setting a lower threshold for treating stimuli as humans when they share their political perspectives.
- Users who are more concerned about social bots and have greater social media experience are willing to pay more for aid in detecting social bots.
- Willingness to pay for aid in detecting social bots varies based on perceived sensitivity and criterion.

Ch 5. Conclusion

Social Bot Detection as a Signal Detection Task

We employed a signal detection theory (SDT) methodology to assess individuals' behavior when responding to tweets that might reflect social bots rather than humans. SDT allows estimating both individuals' *sensitivity*, or ability to distinguish signal from noise and their *criterion* (or decision rule) when treating an uncertain stimulus as a 'signal' or 'noise' (Green, & Swets, 1966; McNicol, 2005). Across three studies involving roughly 2,000 participants, we find that individuals have modest sensitivity to social bots and a criterion that reflects hesitancy to label human personas as social bots.

Social Bot Detection Performance

In study 1, we asked lay participants whether each of 50 Twitter personas was a human or social bot. We used the agreement of two machine learning models to estimate the probability of each persona being a bot and then used those estimates to evaluate human performance. On average, participants had modest sensitivity (d') and a criterion that favored responding 'human.' Exploratory analyses found greater sensitivity for participants (a) with less self-reported social media experience, (b) greater analytical reasoning ability, and (c) who were evaluating personas with opposing political views. Some patterns varied with participants' political identities. These results replicated within the control groups used in studies 2 and 3.

Improving Social Bot Detection

After estimating baseline social bot detection performance and identifying individual attributes that affected it, we investigated means to improve human social bot detection. In study 2, we test the effectiveness of a social bot detection decision support tool similar to a commercially available product. Computer scientists have developed social bot detection tools using Artificial Intelligence (AI) systems for over a decade (Cresci, 2020; Davis, Varol, Ferrara, Flammini & Menczer, 2016; Beskow & Carley, 2018). Some tools require users to leave the Twitter environment to investigate personas (Botometer, Bot Sentinel). We elected to test the utility of decision support overlays that show a bot indicator signal next to a persona's user name within Twitter. (Kats, 2022).

With Aid. In study 2, we compared the social bot detection performance of participants shown the bot indicator score (aid group) with that of a control group and group given a reminder to “look for bot cues.” The bot indicator displayed either a human or robot icon and bot probability score on a persona’s profile near their name. When provided aid, participants had improved sensitivity relative to a control group, and the group provided a reminder. Participants who were provided aid also had less conservative response thresholds. When asked if they used these aids to guide their judgments, participants responded positively, and they had less overconfidence than the control and reminder group participants.

These findings correspond to results that demonstrate individuals are generally willing to accept and make use of decision support tools if they trust them and the consequences of the decision are relatively low (Glikson, & Woolley, 2020; Hoff, & Bashir, 2015; Lee, & See, 2004; Pavlou, 2003; Venkatesh, Thong, & Xu, 2016; Kluttz, & Mulligan, 2019; Siau & Wang, 2018).

With Training. In study 3, we compared the social bot detection performance of participants provided training to that of a control group and a group given the same aid as in study 2. Participants who received social bot detection training had improved sensitivity. However, they remained hesitant to respond ‘bot.’ Our results demonstrated individuals are capable of being trained to detect cues associated with deception (Hauch, Sporer, Michael, & Meissner, 2016; Hartwig, Granhag, Strömwall, & Kronkvist, 2006) using a protocol focused on orienting individuals to signals of representativeness (Mohan et al., 2017; Mohan et al., 2014).

The training protocol developed in this study first provided a framework of representativeness about what distinguished social bots from average human social media users. Like that of Mahan et al. (2014, 2017), this finding reveals the benefit of orienting trainees to relevant signals to guide their search and judgment. In this case, the relevant cues were associated with the objectives of social bot developers: amplifying narratives to a broad audience. Heuristic training benefits users when algorithms are unavailable. As adversarial social bot developers seek to circumvent new approaches to detection, these heuristics may need regular updating. Correspondingly, continued research to identify these signals and update social bot training protocols would be an essential area to focus social bot mitigation efforts.

Bot Responses Reduced Probability of Sharing

After judging personas as either a human or a bot and rating their confidence in that response, we asked participants to respond if they would be willing to retweet content from that persona if they agreed with its content. On average, regardless of test condition, participants would only retweet content from 15% of personas. Participants who responded ‘bot’ on a given social bot detection trial had a probability near 0% of retweeting. This may reflect individuals’ wish to avoid being complicit in fake personas’ behavior or losing their online epistemic reputation (Altay, Hacquin, & Mercier, 2020a, 2020b).

Social bots are linked to increasing low credibility information (Vosoughi, Roy, & Aral, 2018). Improving individuals’ ability to detect and label bots can reduce their willingness to share bot content. Therefore, both of our interventions have the potential to help combat the spread of false narratives by helping prevent Twitter users from being duped by social bots.

Factors Affecting Social Bot Detection

Myside Bias. Across all three studies, we found effects of myside bias. In studies 1, 2, and 3, regardless of test conditions, myside bias led participants to lower their thresholds for responding ‘bot’ when viewing personas of opposing views. Furthermore, an asymmetric difference between liberals and conservatives appeared across all three studies. While conservatives had higher probabilities of responding ‘bot’ when viewing liberals, liberals had much higher probabilities. Conservatives’ criterion shift did not impair their sensitivity, but it did liberals.

The effects of myside bias on sensitivity also proved asymmetric. Liberals viewing liberals maintained the same relative sensitivity for social bots. However, because they lowered their thresholds to such a large degree when viewing conservatives, they assumed most conservatives were bots and failed to investigate them adequately. Conservatives had little or lower sensitivity when viewing conservatives relative to when they viewed liberals. Conservatives considering liberals had increased sensitivity. Interpreting these findings depends partly on how one considers biases within these groups.

Analytical Reasoning. Across three studies, analytical reasoning did not appear to benefit sensitivity for social bot detection. The results of previous studies differ regarding the relationship between cognitive skills and the ability to detect deceptive information. Pennycook and colleagues have found that people with higher CRT scores have more ability (Pennycook & Rand, 2019; Bronstein, Pennycook, Bear, Rand & Cannon, 2019; Ross, Rand & Pennycook, 2021). Here, we did not observe the same benefits.

Though ancillary to the central aims of this research, the consistent trends we observed involving analytical reasoning support arguments made by others that myside bias reflects analytical reasoning deployed to find confirmational evidence in favor of one's previous beliefs (Stanovich, 2022; Stanovich, & West, 2007; Stanovich, West, & Toplak, 2013). We observed analytical reasoning interacting with myside bias when participants viewed personas of opposing views both to exacerbate the lowering of thresholds or enhance discernment. We never saw it interact to reduce those effects. Instead, like others (Kahan, 2012; Drummond & Fischhoff, 2017; Haidt, 2012; Stanovich & West, 2007; Strickland, Taber & Lodge, 2011), we observed participants who scored highest in cognitive reflection displayed the most ideologically motivated cognition.

Social Media Experience.

We expected that social media experience would benefit participants' ability to detect social bots; it either did not or impaired their performance. We initially believed social media experience would foster increased expertise in detecting social bots. Our studies' measure focused on the length of experience within social media, average use, and engagement estimates. In study 1, we hypothesized individuals with high social media experience would likely have numerous interactions with 'typical' human and 'typical' bot profile features and behaviors. If these experiences yielded domain-specific expertise, they would have directed their attention toward relevant social bot signals.

Studies typically find that experts outperform novices in discrimination tasks (Allen, Mcgeorge, Pearson & Milne, 2004; Bond, 2008; Spengler, White, Ægisdóttir, Maugherman, Anderson, Cook & Rush, 2009; Cañal-Bruland & Schmidt, 2009). Experts gain increased situational awareness for their tasks by directing their search more effectively than novices. Experts focus on information sources most relevant to their analytical objective (Langley, 1985;

Weiss, & Shanteau, 2003; Landy, 2018; Endsley, 2018). However, as others have found, we observed that experience does not confer expertise (Ericsson, 2018; Bisseret, 1981; Rikers, Schmidt & Boshuizen, 2000; Witteman & Tollenaar, 2012). Without meaningful feedback to prove them wrong frequent social media users may have learned the wrong detecting heuristic.

Individuals' Willingness to Pay Adopt Social Bot Detection Aids

Following the social bot detection task, we asked participants to rate how much they would be willing to pay for a monthly social bot detection service. They could enter any amount. Though exploratory, our results concerning individuals' willingness to pay for aid in detecting social bots offer a few insights. Our findings suggest a metacognitive influence of each group's perceptions of their performance on their willingness to pay.

Control group. In studies 2 and 3, participants with higher sensitivity and stringent criteria were less willing to pay for a social bot detection service than participants with low sensitivity and lenient response thresholds. Those quick to respond 'bot' who performed poorly were willing to pay for assistance. This suggests that these participants' willingness to pay for aid was correlated with their metacognitive reflection on their detection performance.

Aid group. Participants in studies 2 and 3, who had access to bot indicator scores as they made their judgments had varied willingness to pay responses. In both studies, participants with more conservative thresholds resulted had an increased willingness to pay more for aid. This suggests that as they encountered instances where the bot indicator score was above the 50% threshold, they were challenged to respond 'bot.' Perhaps this form of feedback convinced them that aid might be worth paying for.

In study 2, increased sensitivity, reflecting a reliance upon the assistance during the study, was associated with a willingness to pay more for a service. In study 3, higher sensitivity was modestly negatively correlated with willingness to pay. The only difference between studies 2 and 3 that may explain this result was the short social bot orientation video shown to study 3 participants: this video described social bots and the challenges in detecting them. Given the inconsistency in these findings, we recommend further investigation of these phenomena before any definitive conclusions are made.

Training group. Training group Participants provided training, who had higher sensitivity, were willing to pay more for a social bot detection than those with lower sensitivity. Higher sensitivity due to training reflects participants' appropriate use of the trained decision rules. We speculate that participants who worked diligently at the task and thus performed well recognized either their lack of prior abilities and hence need for assistance or reflected on the cognitive effort required to conduct the task properly. They were thus more willing to outsource that mental effort. As these are speculations, future research may wish to investigate these hypotheses.

Social Bot Concerns. Participants were consistently more worried about social bots influencing others than they were concerned about being affected by social bots themselves. This finding corresponds to the large body of work regarding asymmetries between self-perception and social perception of other people's actions and judgments (e.g., Pronin, Gilovich, & Ross, 2004). The difference between how individuals view themselves versus others may have led participants to attach greater faith to their own ability to avoid being duped relative to others (e.g., Dunning, Meyerowitz, & Holzberg, 1989).

Social Media Experience. In studies 2 and 3, higher scores on the social media experience survey corresponded to a willingness to pay more for a social bot detection service. The social media experience survey used throughout these studies measures the length of use, levels of engagement, and how frequently users are on different social media platforms. A posthoc examination of study 2 and 3 responses from this survey that had the strongest positive correlation with a willingness to pay were frequency of use and level of engagement compared to users' number of accounts and years of use (see Appendix F). This suggests that the variance in willingness to pay accounted for by the broader measure of social media experience may be more closely linked to one's level of engagement.

Limitations

Normative Social Bot Detection Estimates.

Our conclusions depend upon the accuracy of the normative bot indicator scores provided by three machine learning systems: Bot-hunter, Botometer, and Bot Sight. These models may have been trained on unrepresentative and mislabeled training sets, with unknown effects on our results. Additionally, incomplete information can lead to faulty machine learning models. In comparing the predictions of each of these models in non-selected stimuli, we observed consistent differences between their predictions. We suspect this was due to how each model was trained and how each developer selected their 'bot' criteria.

Our results depend on the precision of the estimates provided by these bot detection algorithms. We caution researchers interested in applying our techniques in future studies to carefully consider why each model produces different predictions and develop reasonable means of confirming results across models as you look for normative estimates. While we believe we made the best use of available social bot detection tools to estimate normative bot signal strengths, our statistical inferences may still be lacking and require further refinement.

Naturalistic Setting.

A second potential limitation of our research was our experimental task. As with other simulated experiences (Wald, Khoshgoftaar, Napolitano, & Sumner, 2013; Aiello, Deplano, Schifanella, & Ruffo, 2012), the validity of our task depends on how well it evoked real-world behavior. We used actual personas and set a pace akin to the rapid evaluation of everyday Twitter use. However, we did not provide access to the persona profile pages that suspicious users might examine. Nor did we allow participants to scroll through a users' collection of Tweets or explore a persona's social network. Hence, we might have underestimated their abilities.

We also may have overestimated their bot detection skill. As most social media users likely operate in a distracted state and rarely are asked to pause and investigate profiles, our setting may have induced an unnatural state of vigilance. At best, we think one could interpret our findings as suggesting that most perform poorly even when alerted to social bots. Therefore, in the real world, with more demands on attention, most people are likely being supped by social bots.

Estimate of Social Media Experience.

Additional criticism of our study's lack of findings concerning social media experience may pertain to the survey used across these studies. This survey, adapted from Hou (2017), focuses on two components of online social media use, (1) relative activity levels and (2) level of engagement. This survey asks participants to evaluate these factors across multiple platforms, including Twitter. To perhaps provide more insight into why experience may not help or limit one's ability to detect social bots, future research could investigate a measure more specific to Twitter users.

Future Work

Trust and Adoption of Training and Aids.

Several questions remain regarding the long-term benefits of both interventions tested in our research. We observed individuals using the social bot detection aid to improve their detection sensitivity and criterion accuracy. However, based on their performance, either individuals did not entirely trust the indicator, or some failed to use it properly. Appendix I provides a series of post-hoc figures demonstrating a hesitancy to go with the bot indicator choice around the 50% threshold. However, near the extremes of 99% human and 99% bot, participants did not always defer, instead relying upon their judgment over the bot indicator score. As such, questions remain as to why some participants responded differently.

As the factors that limited sensitivity, namely myside bias and analytical reasoning, did not interact by test condition, open questions also remain regarding why some chose to trust their judgment over the bot indicator. Additionally, there is reason to believe that repeated exposure to automated decision support tools may calcify or dilute one's trust in them depending on the congruence of their recommendations with one's views (Glikson, & Woolley, 2020; Lee & See, 2004; Siau, & Wang, 2018). Therefore, a long-term examination of user adoption of social bot aids may help identify why individuals lose trust in the output of these systems.

A similar long-term assessment of the benefits of training on social bot detection is warranted. Before any policymakers or organizations consider an investment in the benefits of training, they should consider how persistent these benefits are. Like other vigilance training, teaching skills to improve vigilance may erode over time. They also may become obsolete as social bot developers modify their approach to avoid detection.

Aside from assessing the long-term benefits of training, future research should investigate the effects of training interventions focused not only on sensitivity but also on one's criterion, such as training that leverages high base rates (Kaivanto, 2014; Kaivanto, Kroll, & Zabinski, 2014; Wolfe, Brunelli, Rubinstein, & Horowitz, 2013). Additionally, as myside bias appeared to affect the criterion of both liberals and conservatives, training designed to reduce these effects may also improve social bot detection.

Policy Implications and Recommendations

Social bots take advantage of our limited attention affecting what information we search for, understand, remember and share. Falling prey to social bot personas exposes individuals to misleading content and an increased likelihood of unwittingly sharing bot propagated content, including misinformation and disinformation, with others. Policy-makers should consider our findings, and other social and computer scientists expand upon them to help reduce the risks social bots pose to societies.

To protect users, platforms may either provide bot indicator aids or training. Unfortunately, the incentives to alert users to social bots and take action to remove them run counter to the social media platforms' economic incentives. Their business models rely upon a count of total users and their activity levels. Social bots increase both.

The problem of social bots has reached a level of concern where actions to reduce social bot deception are needed. To protect social media users, the commercial sector could deliver these services at a cost to meet their demand. Additionally, organizations may wish to invest in these solutions to protect their members or as a service to society. Finally, governments could mandate new regulations to require social media platforms to offer these means to protect users or subsidize their delivery.

Our findings establish (1) the baseline ability of individuals to detect social bots, (2) the benefit of real-time bot indicator aids, and (3) an effective method of social bot detection training. Moreover, (4) the more individuals recognize personas as social bots, the less likely they are to share their content. We believe our findings demonstrate credible evidence of two methods, aid, and training, that can help protect social media users from social bot deception. By reducing deception, we can reduce the spread of low credibility information by social bots.

Appendix A

Stimulus Selection and Normative Algorithms

Selecting stimuli across studies has involved a process of determining and then confirming social bot probabilities. We refer to these probabilities as bot indicator scores throughout our studies. The following sections describe the social bot detection algorithms used to estimate the normative bot indicator scores as well as our process of selecting appropriate stimuli for our studies.

Study 1

Study 1 used Bot Hunter's Tier 1 model (Beskow and Carley, 2018) and Botometer (Botometer, 2021) to estimate our normative bot indicator scores. Approximately 4,500 Twitter personas, particularly active during the 2018 U.S. midterm election were evaluated by both algorithms. Subsequently, 100 bins of personas were ordered by Bot Hunter bot probabilities (i.e., 1%, 2%, … 98%, 99%). Within each of these bins, we evaluated Botometer bot probabilities, and randomly selected personas where concordance was within 0.1%. We lessened this criterion for a handful of personas with Bot Hunter bot probabilities greater than 85% where Botometer scores were above 50%, but not as high.

The following figures provide details regarding the features used by Bot Hunter's models as well as the overall concordance of these models' predictions for the 4,500 personas evaluated.

Figure 1A

Table of Bot Hunter Model Features

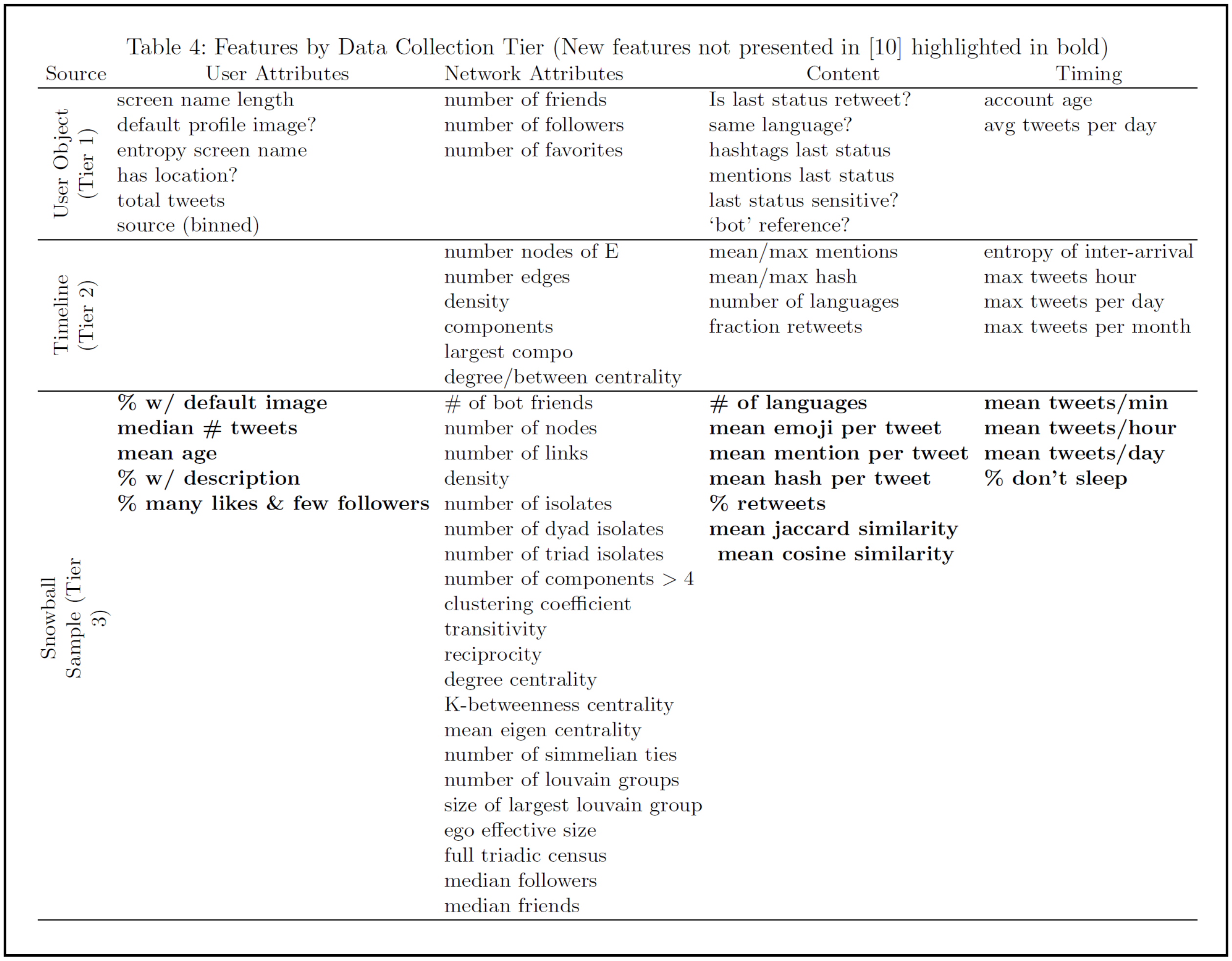


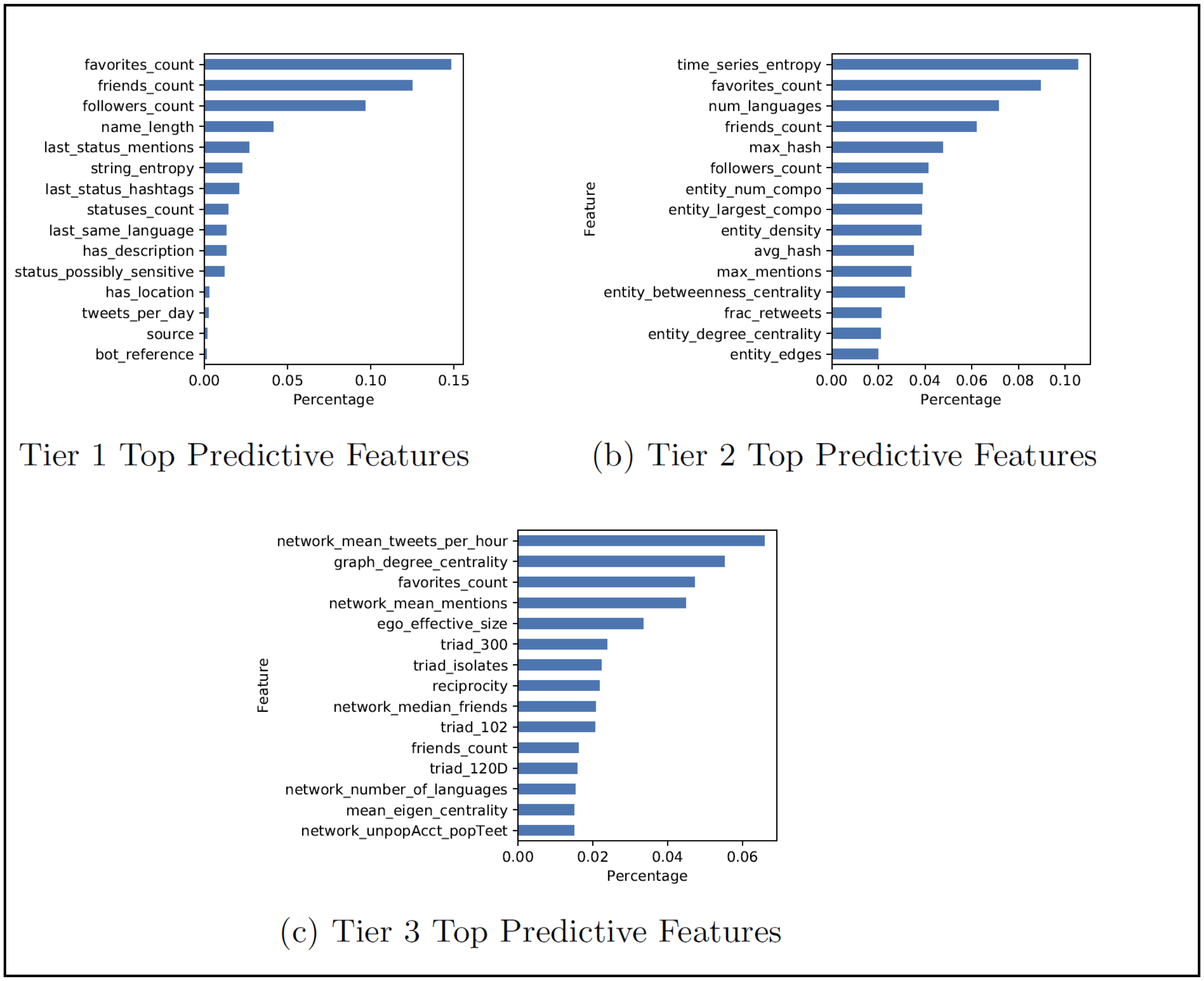
Table 4: Features by Data Collection Tier (New features not presented in [10] highlighted in bold)

Source	User Attributes	Network Attributes	Content	Timing
User Object (Tier 1)	screen name length default profile image? entropy screen name has location? total tweets source (binned)	number of friends number of followers number of favorites	Is last status retweet? same language? hashtags last status mentions last status last status sensitive? 'bot' reference?	account age avg tweets per day
Timeline (Tier 2)		number nodes of E number edges density components largest compo degree/between centrality	mean/max mentions mean/max hash number of languages fraction retweets	entropy of inter-arrival max tweets hour max tweets per day max tweets per month
Snowball Sample (Tier 3)	**% w/ default image** **median # tweets** **mean age** **% w/ description** **% many likes & few followers**	# of bot friends number of nodes number of links density number of isolates number of dyad isolates number of triad isolates number of components > 4 clustering coefficient transitivity reciprocity degree centrality K-betweenness centrality mean eigen centrality number of simmelian ties number of louvain groups size of largest louvain group ego effective size full triadic census median followers median friends	**# of languages** **mean emoji per tweet** **mean mention per tweet** **mean hash per tweet** **% retweets** **mean jaccard similarity** **mean cosine similarity**	**mean tweets/min** **mean tweets/hour** **mean tweets/day** **% don't sleep**

Note. Source: Beskow and Carley, 2018.

Figure 2A

Bot Hunter Model Feature Predictive Importance using Random Forest Algorithm



Tier 1 Top Predictive Features

(b) Tier 2 Top Predictive Features

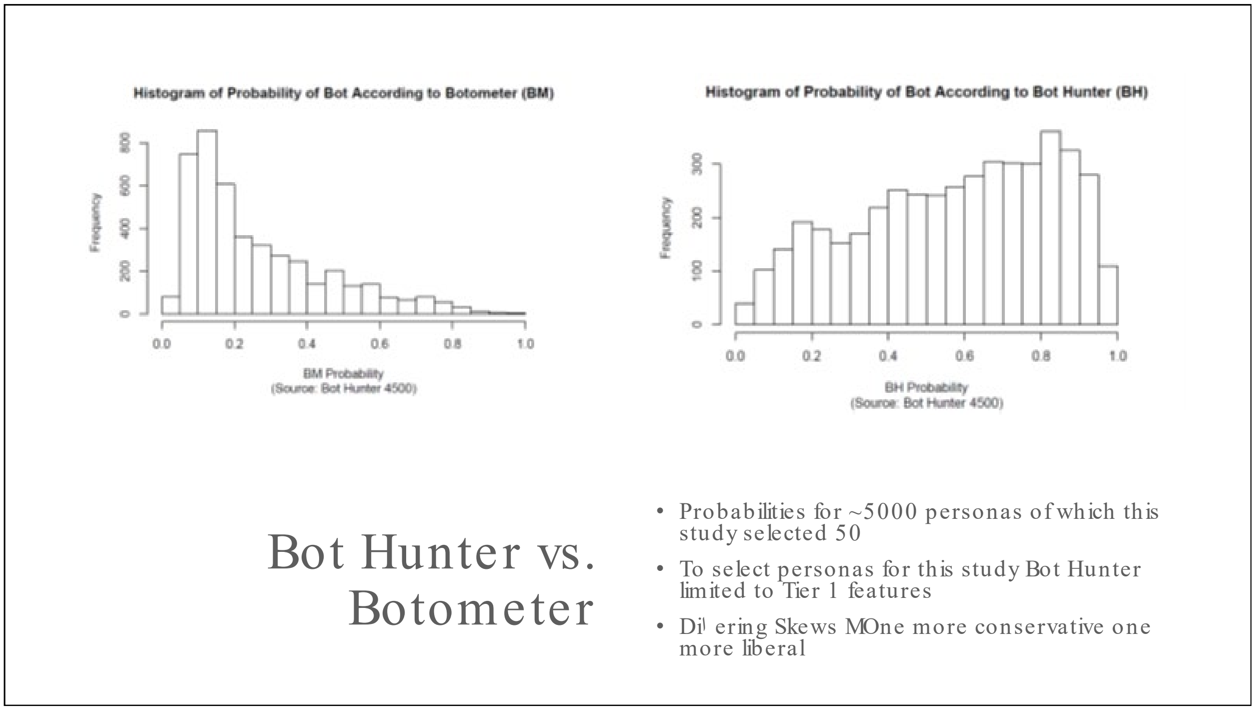
(c) Tier 3 Top Predictive Features

Note. Source: Beskow and Carley, 2018.

After selecting 100 personas, of which we only used 50, we compared the relationship between Bot Hunter Tier 1 bot probabilities and Botometer's probabilities. They both had varying skews in the distributions of probabilities produced and the correlation between probabilities was low.

Figure 3A

Bot Hunter Botometer Prediction Comparison for ~4500 personas



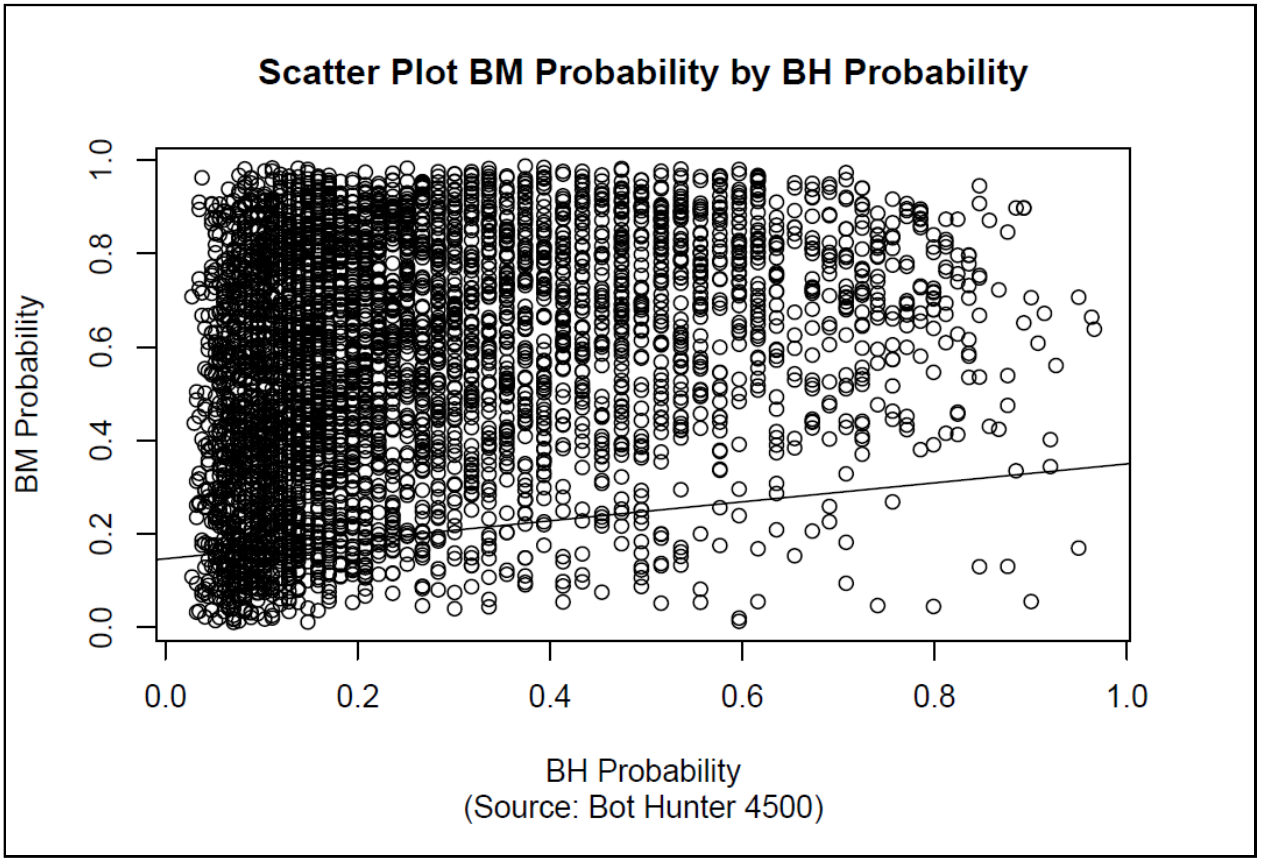
Bot Hunter vs. Botometer

- Probabilities for ~5000 personas of which this study selected 50
- To select personas for this study Bot Hunter limited to Tier 1 features
- Differing Skews MOne more conservative one more liberal

Note. Bot Hunter's Tier 1 model and Botometer provided probabilities for a set of Twitter personas independently. The distribution of each of the model's outputs varied in their skew.

Figure 4A

Scatter Plot of Bot Hunter and Botometer predictions for ~4500 personas

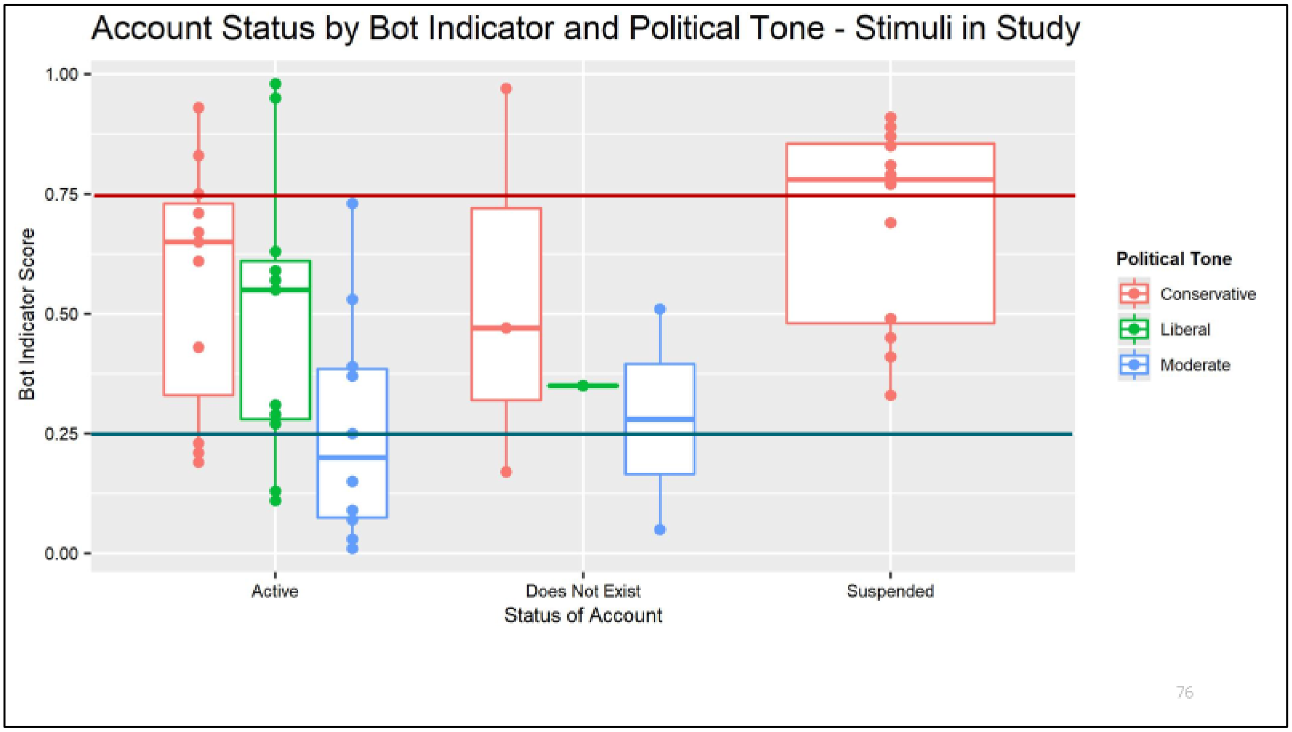


Note. This is a plot of Bot Hunter's and Botometers' predictions for the ~5000 personas shown in the histograms in Figure 3. The models had a correlation of 0.27 for their predicted bot probabilities.

As part of study 1's validation of stimulus selection, we investigated the personas used. We looked for indicators that personas were validated as probable bots or likely humans. If an account was suspended, it was likely due to being a bot, or acting in a manner contrary to Twitter's guidelines. This analysis took place after the January 6th, 2020 riots; a time when Twitter took extreme measures to remove problematic accounts.

Figure 5A

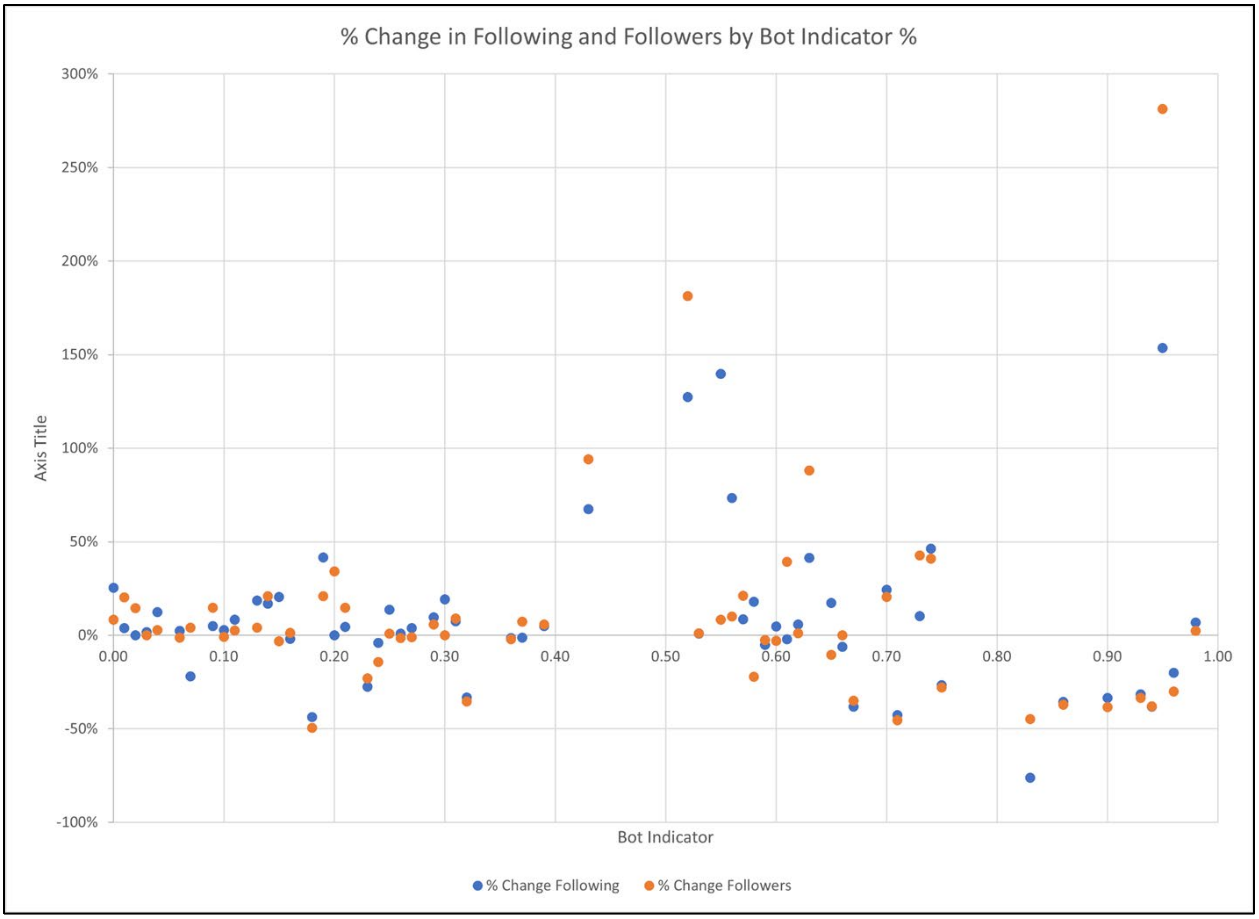
Findings from Post Study 1 Analysis of Accounts – Account Status



Note. Nearly a year after study 1 ended, we investigated the 50 personas used in this study. We found that many of the accounts with bot probabilities greater than 75% had been suspended. Many of those with bot probabilities less than 25% were still active.

Figure 6A

Findings from Post Study 1 Analysis of Accounts – Change in Social Network



Note. While investigating the account status of the personas used in study 1, we also logged changes in their following and followers count. We found that as the bot indicator score increased, accounts that had not been suspended, but were probable bots, had significant changes to the size of their social networks.

Study 2

Unlike study 1, where we used a set of stimuli used by Bot Hunter's developers to train and test their models, we began our search for appropriate stimuli by working through online Twitter personas using NortonLifeLock's Bot Sight tool. Process:

1. I began my search by looking for confirmed bots from existing bot detection training sites provided by Botometer (https://botometer.osome.iu.edu/bot-repository/)
2. In the initial collection phase, I selected personas randomly across the full range of bot probabilities.
3. I tried to find at least one persona within every 5% bin (i.e., 0%-5%, 5%-10%...90%-95%, 95%-100%).
4. Every selection was then rated by Botometer. Stimuli were only selected when concordance appeared between the Bot Sight ratings, and Botometer's ratings. As there was rarely perfect concordance, any stimuli that was deemed likely a bot, or a human by Bot Sight, and also likely a bot, or a human by Botometer was included.
5. After collecting 30-40 personas in this manner, I then focused on finding personas for bins that had not yet been filled. This process was particularly challenging around the 50% threshold.
6. After I full set of personas were filled, with 20 liberal personas, 20 moderates, and 20 conservative personas, I then ran those personas through Bot Hunter's algorithm to determine their Bot Hunter probabilities.
7. If Bot Hunter's estimates did not match Bot Sight's and Botometers, those personas were replaced by returning back to steps1 and 2.
8. Collecting the full set of stimuli in this manner took a few days.
9. Bot Hunter's scores served as the final normative estimate for our stimuli.

The following figures depict the final bot estimates from the three separate bot probability models (Bot Hunter, Botometer, and Botsight) and their relationships with the assessed political tone of the given stimuli (liberal, moderate, conservative). Moderate stimuli do not have any overt political messages, themes, etc.

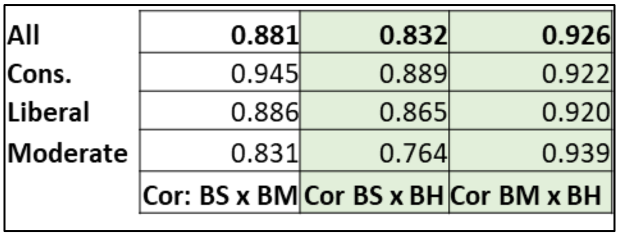
Bot Hunter's scores are based on a limited set of features a typical user would encounter on a user's profile. As such, it serves as the 'best' normative score we might expect users to match. The other two models include additional features and are used to corroborate the bot hunter scores.

To determine confirmation of these algorithmic predictions, my central two concerns that drove stimulus selection were:

1. Uniform distribution of Bot Hunter scores across political tones
2. Concordance in predictions linearly, and concordance across the 50% Bot Hunter probability threshold. Even if a prediction diverged in its probability, I wished to avoid one model predicting 'bot' as another predicted 'human'.

Table 1A

Correlations Between Bot Detection Algorithm Predictions



All	**0.881**	**0.832**	**0.926**
Cons.	0.945	0.889	0.922
Liberal	0.886	0.865	0.920
Moderate	0.831	0.764	0.939
	Cor: BS x BM	**Cor BS x BH**	**Cor BM x BH**

Note. BS stands for Bot Sight, BH stands for Bot Hunter, BM stands for Botometer

The correlation between Bot Hunter and Botometer was 0.926. The correlation between Bot Hunter and Botometer was 0.88 and between Botometer and Bot Sight was 0.83. The proportion above/below 50% according to Bot Hunter is balanced across political tone conditions.

Figure 7A shows average Bot Hunter probabilities for the 20 conservative, 20 liberal, and 20 moderate stimuli. Figures 8, 9, and 10 demonstrate the concordance of scores across the three bot detection algorithms.

Figure 7A

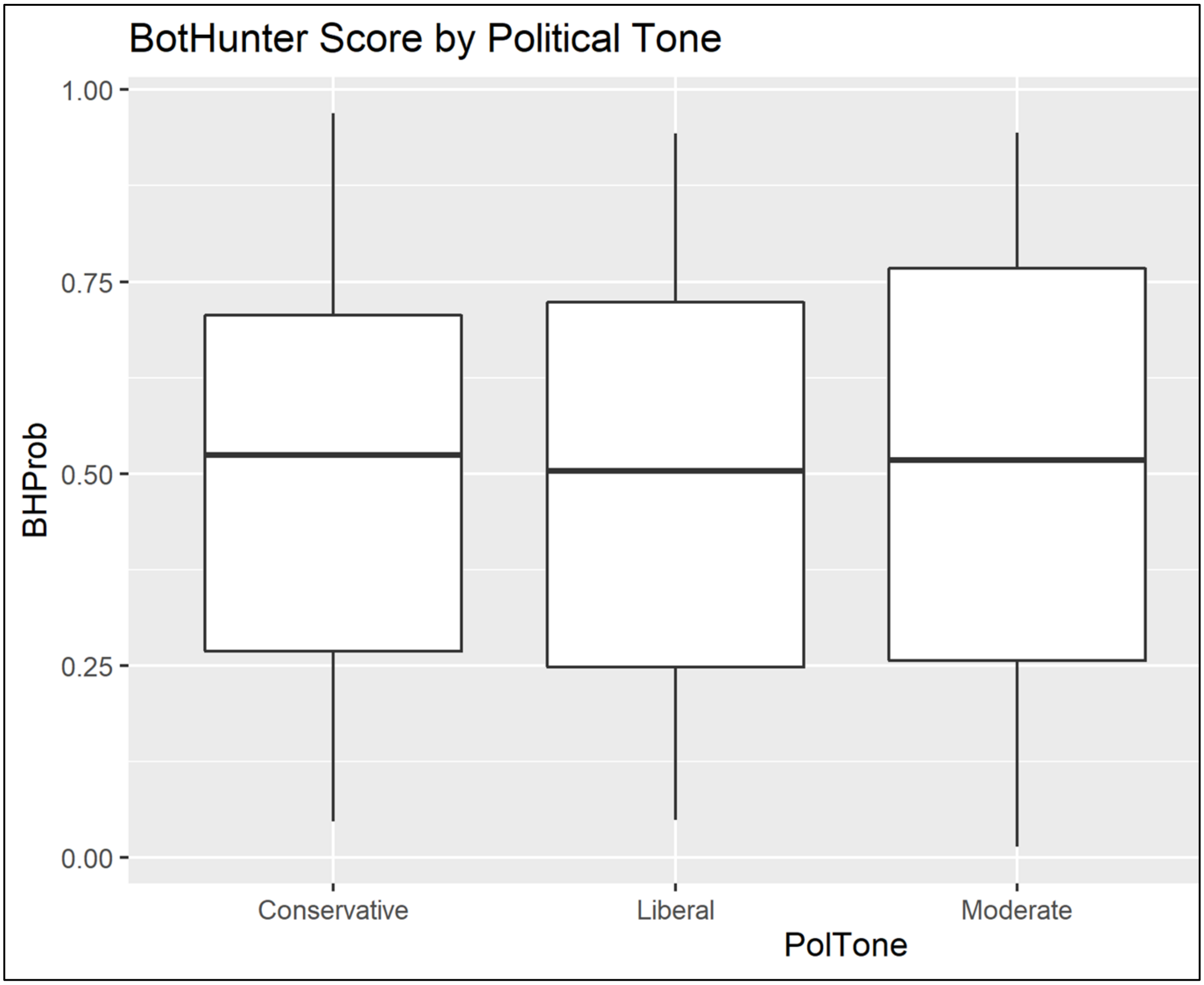

BotHunter Score by Political Tone
1.00
0.75
0.50
0.25
0.00
BHProb
Conservative
Liberal
Moderate
PolTone

Figure 8A

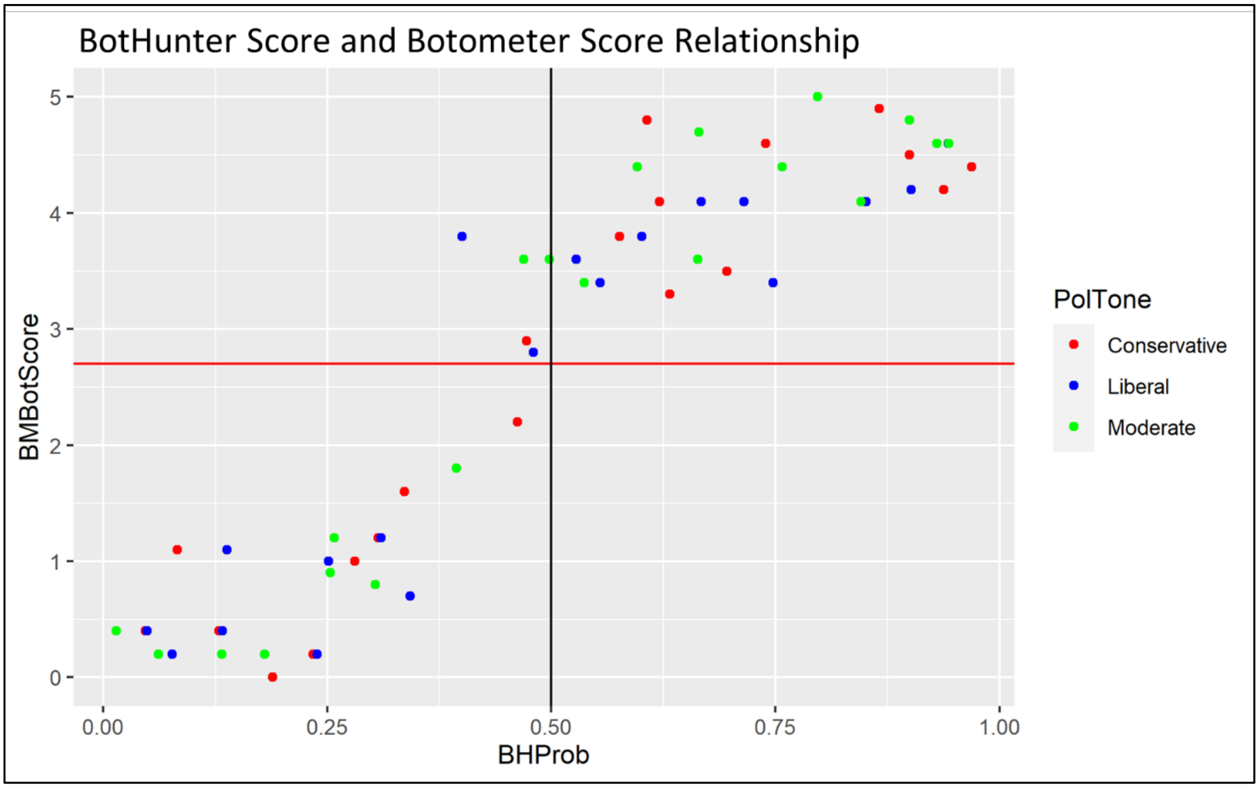


Figure 9A

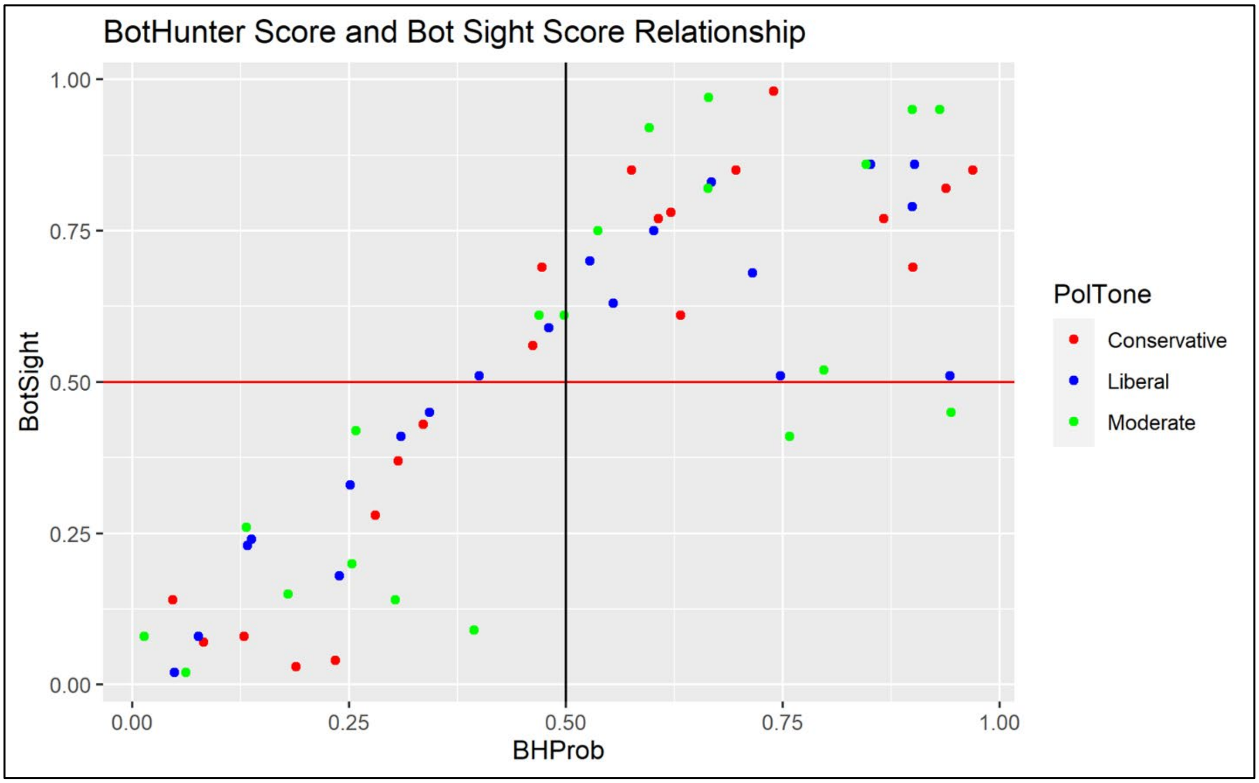
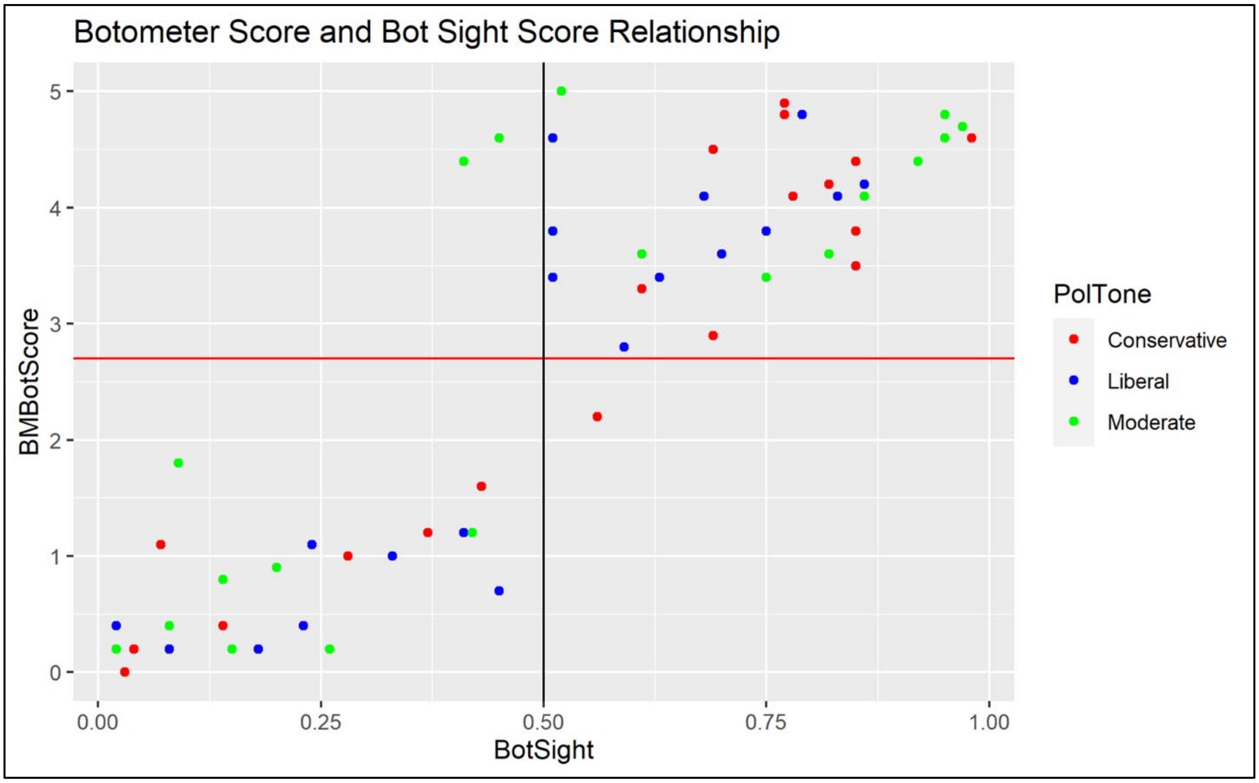


Figure 10A


Botometer Score and Bot Sight Score Relationship
BMBotScore
0
1
2
3
4
5
BotSight
0.00
0.25
0.50
0.75
1.00
PolTone
Conservative
Liberal
Moderate

Appendix B

Social Media Experience Survey

The following survey was adapted from Hou (2017). Social media experience scores were computed by summing across responses for a total score. In study one, questions about Instagram and TikTok were not included. In studies 3, they were. The scoring method is described for each question below. Additionally, descriptive statistics for responses across these studies are provided.

1. Do have an account with any of the following four social media sites (Facebook, LinkedIn, Twitter, or YouTube)? Please select all that apply.

□ Facebook

□ LinkedIn

□ Twitter

□ YouTube

□ Instagram

□ TikTok

□ Other

For every account selected, the score for this response increased by one. A maximum score for study 1 was five, for studies 2 and 3 was seven.

2. How often do you use these different social media sites?

	Never	Rare	Monthly	A few times per month	Weekly	A few times per week	Daily
Facebook	○	○	○	○	○	○	○
LinkedIn	○	○	○	○	○	○	○
Twitter	○	○	○	○	○	○	○
YouTube	○	○	○	○	○	○	○
Instagram	○	○	○	○	○	○	○
TikTok	○	○	○	○	○	○	○
Other	○	○	○	○	○	○	○

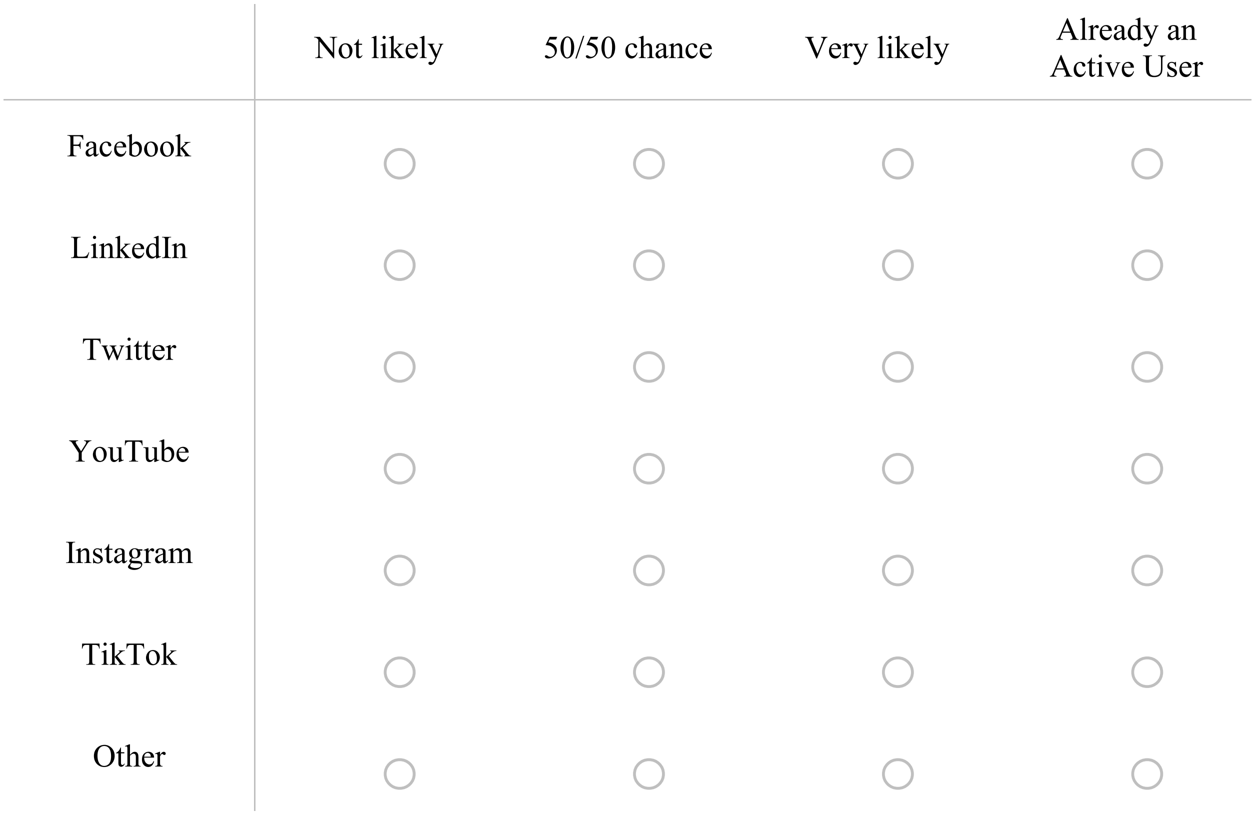
‘Never’ responses = 0, ‘Rare’ responses = 1, ‘Monthly’ responses = 2, ‘A few times per month’ responses = 3, ‘Weekly’ responses = 4, ‘A few times per week’ responses = 5, ‘Daily’ responses = 6. A maximum score for study 1 was 30, for studies 2 and 3 was 42.

3. How long have you been using these different social media sites?

	Never	Less than 6 months	6 months to 1 year	1-2 Years	2-3 years	More than 3 years
Facebook	○	○	○	○	○	○
LinkedIn	○	○	○	○	○	○
Twitter	○	○	○	○	○	○
YouTube	○	○	○	○	○	○
Instagram	○	○	○	○	○	○
TikTok	○	○	○	○	○	○
Other	○	○	○	○	○	○

‘Never’ responses = 0, ‘Less than 6 months’ responses = 1, ‘6 months to 1 year’ responses = 2, ‘1-2 years’ responses = 3, ‘2-3 years’ responses = 4, ‘More than 3 years’ responses = 5, A maximum score for study 1 was 25, for studies 2 and 3 was 35.

4. If you do not currently actively use social media (e.g. at least once a month), how likely you are to start (or become more active) within the next 2 years?



	Not likely	50/50 chance	Very likely	Already an Active User
Facebook	○	○	○	○
LinkedIn	○	○	○	○
Twitter	○	○	○	○
YouTube	○	○	○	○
Instagram	○	○	○	○
TikTok	○	○	○	○
Other	○	○	○	○

‘Not Likely’ responses = 0, ‘50/50 chance’ responses = 1, ‘Very likely’ responses = 2, ‘Already an Active User’ responses = 3, A maximum score for study 1 was 15, for studies 2 and 3 was 21.

5. How often do you engage in each of these social media activities.

	Never	Rarely	Sometimes	Half of the Time	Very Often	Always
Read other's posts or tweets or updates	○	○	○	○	○	○
Post own messages or tweets or updates	○	○	○	○	○	○
"Like" others' posts or links, "follow" or "friend" others	○	○	○	○	○	○
Comment on others' posts or tweets or links	○	○	○	○	○	○
Respond to own posts or tweets or links	○	○	○	○	○	○
Share others' posts or links	○	○	○	○	○	○
Share others' photos or (video) links	○	○	○	○	○	○
Share own photos or videos/links	○	○	○	○	○	○

'Never' responses = 0, 'Rarely' responses = 1, 'Sometimes' responses = 2, 'Half of the time' responses = 3, 'Very Often' responses = 4, 'Always' responses = 5, A maximum score for study 1 was 25, for studies 2 and 3 was 35.

6. Have you ever engaged in any of the following activities?

	No	Yes
Created or hosted an event using social media tools	○	○
Created a social media page or group	○	○
Created a YouTube channel	○	○
Wrote a tweet using a hash tag (#)	○	○
Create a branded hashtag challenge	○	○

‘No’ responses = 0, ‘Yes’ responses = 1, A maximum score for study 1 was 4, for studies 2 and 3 was 5. “Created a branded hashtag challenge” was added for studies 2 and 3 for TikTok users.

7. What are your views toward social media?

○ social media is something that I could easily do without

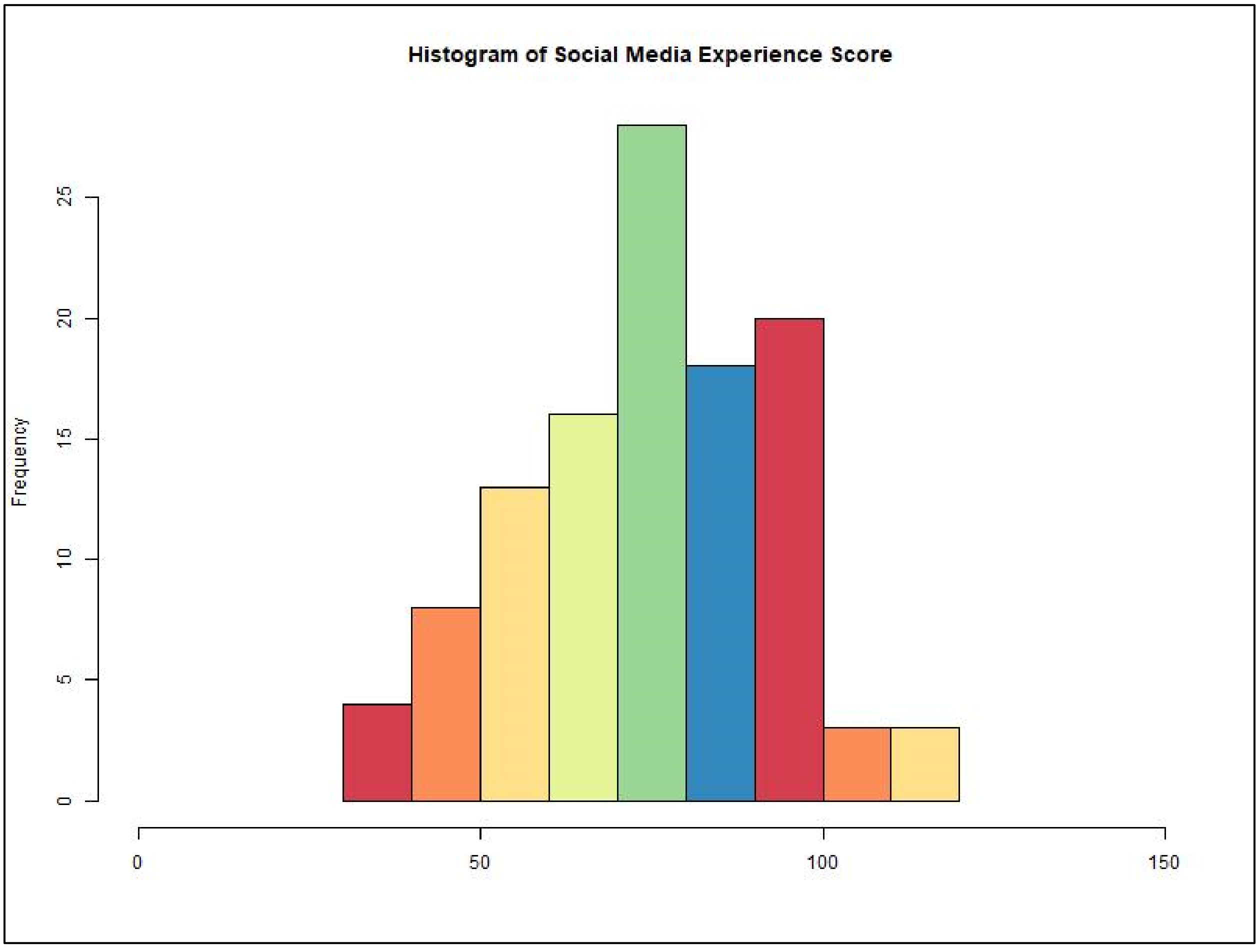
○ social media is enjoyable but not very important

○ social media is an important part of life that I wouldn’t want to live without

‘…do without’ responses = 0, ‘…very important’ responses = 1, “…wouldn’t want to live without” responses = 2. A maximum score for all studies was 2.

Figure 1B

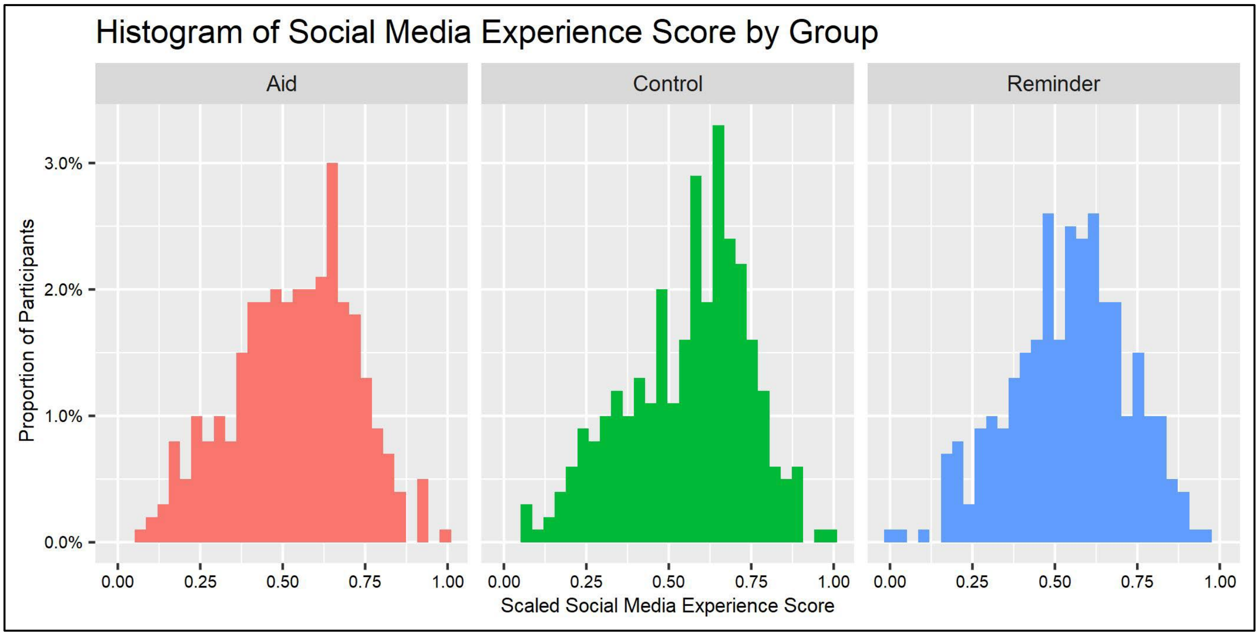
Study 1



Note. N = 113

Figure 2B

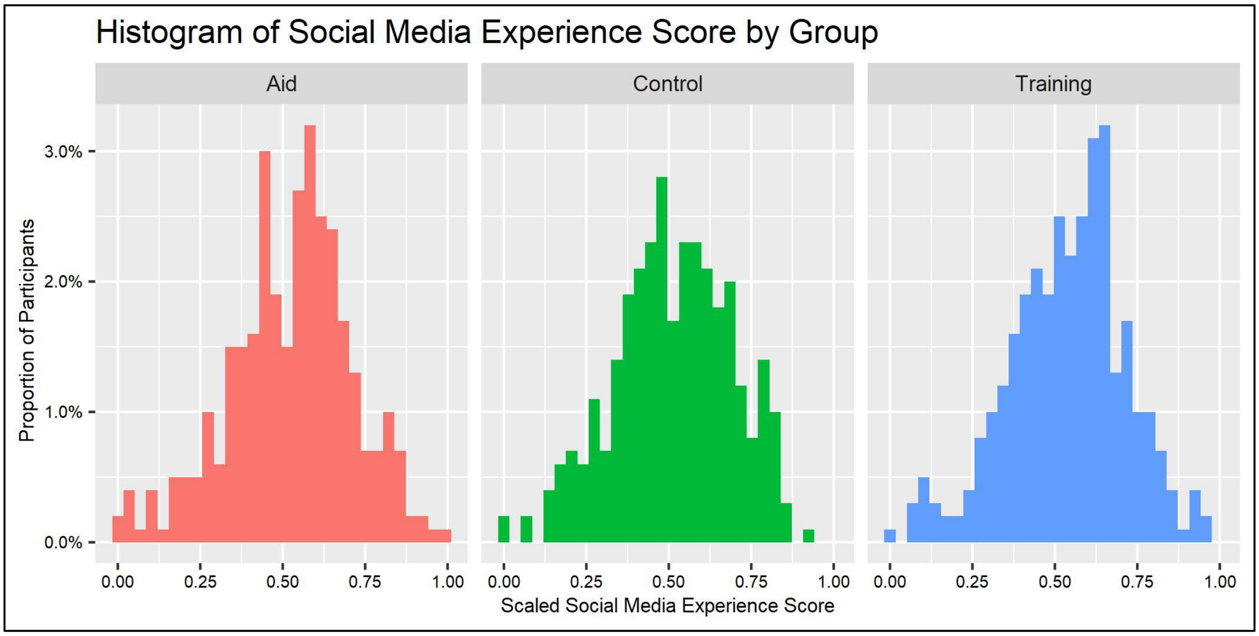
Study 2



Note. N = 924

Figure 3B

Study 3



Note. N = 976

Appendix C

Political Views Survey

In study 1, the following question was asked to determine participants political views.

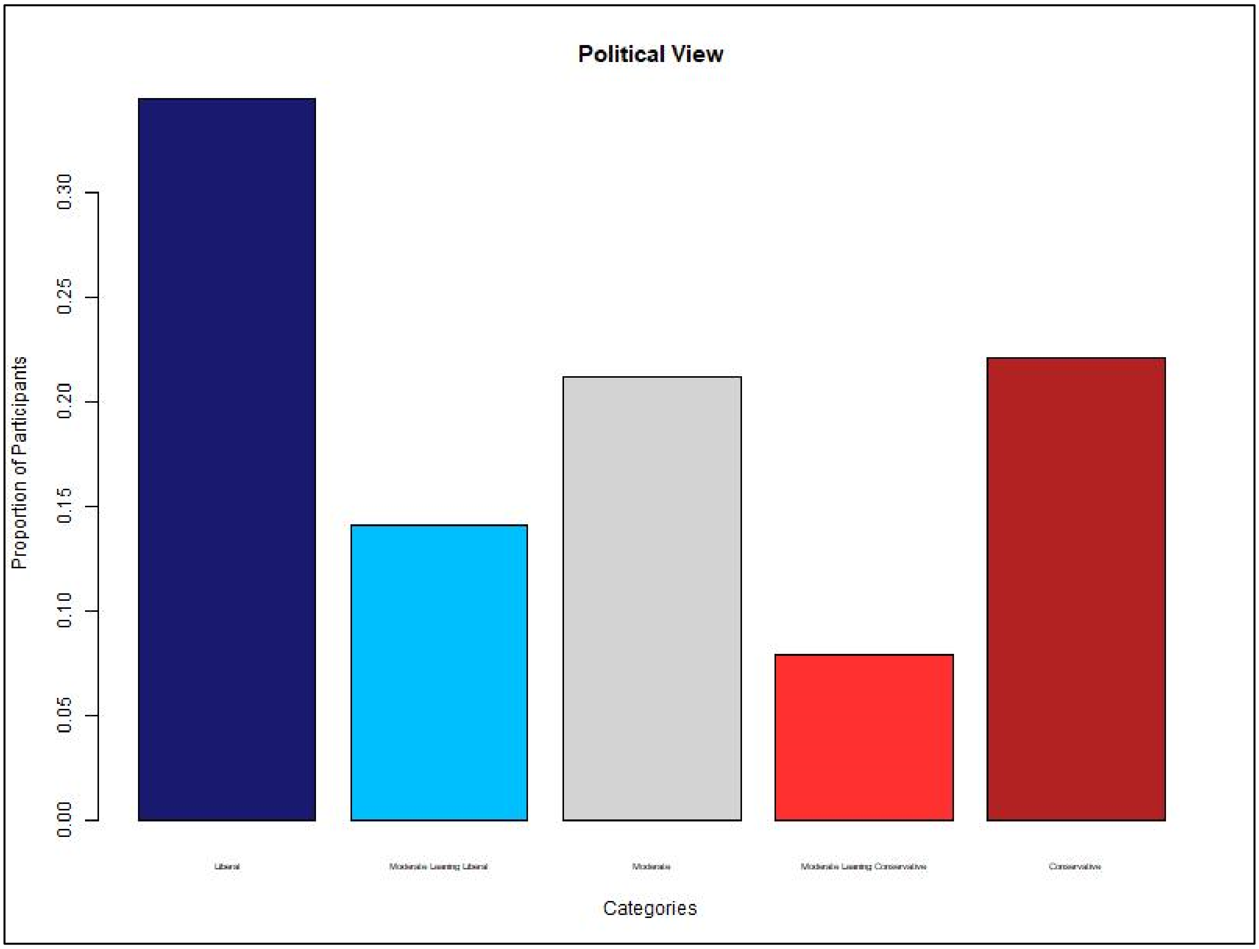
1. When it comes to most political issues, do you think of yourself as a…?
    a. Liberal
    b. Moderate leaning Liberal
    c. Moderate
    d. Moderate leaning Conservative
    e. Conservative
    f. Prefer not to say

In studies 2 and 3, this same question with an extended scale was used.

1. When it comes to most political issues, how do you see yourself?
    a. Strong liberal
    b. Liberal
    c. Moderate leaning liberal
    d. Moderate
    e. Moderate leaning conservative
    f. Conservative
    g. Strong conservative

Figure 1C

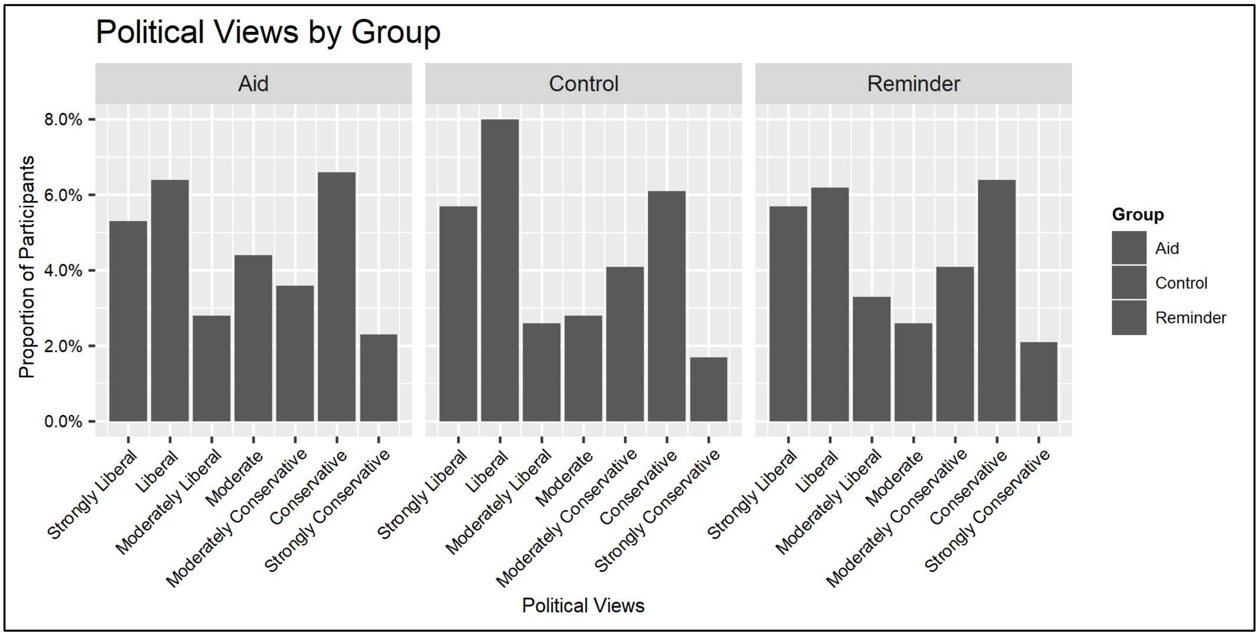
Study 1



Note. N = 113

Figure 2C

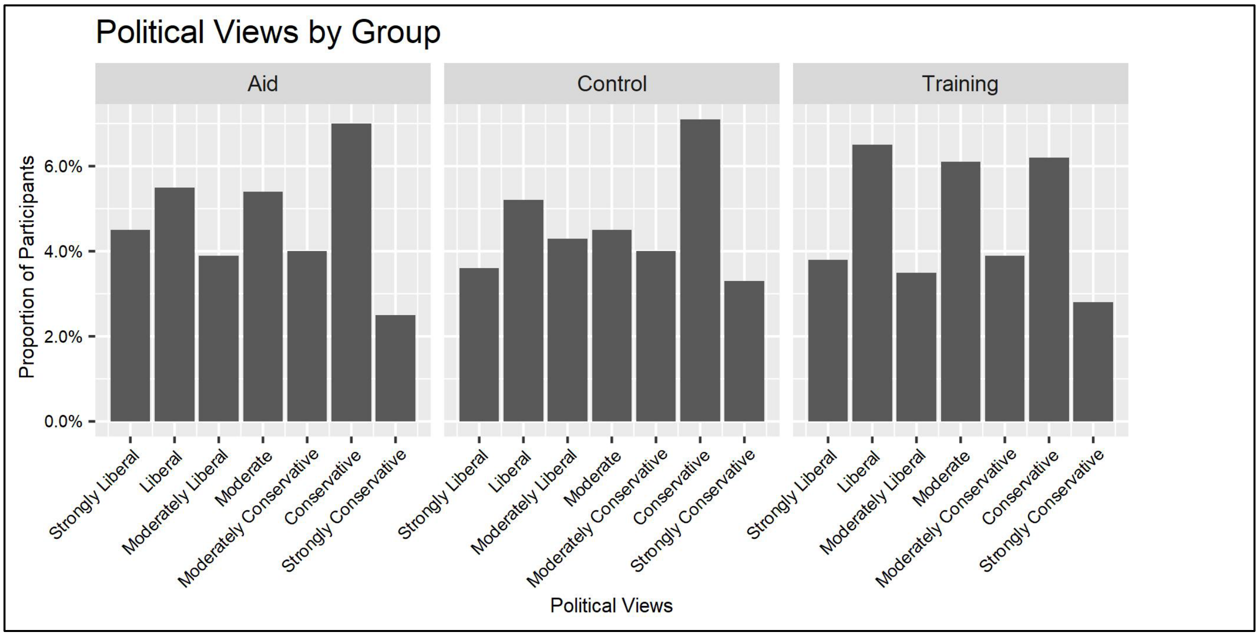
Study 2



Note. N = 924

Figure 3C

Study 3



Note. N = 976

Appendix D

Cognitive Reflection Test

We used the CRT and its variants across these studies. Those questions and descriptive statistics of participants' responses to these questions follow.

Study 1 used the following CRT questions derived from Frederick, 2005.

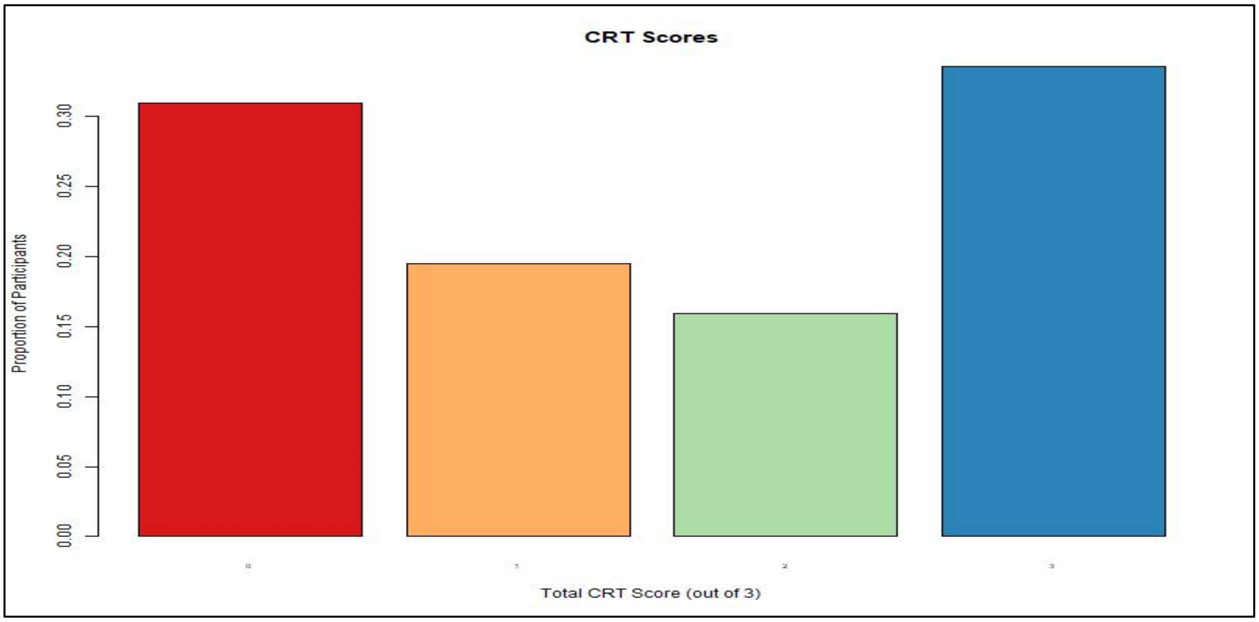
1. A bat and a ball cost $110 in total. The bat costs $100 more than the ball. How much does the ball cost? _____ dollars. [Answer: $5]
2. If it takes 5 machines 5 minutes to make 5 widgets, how long would it take 100 machines to make 100 widgets? _____ mins [Answer: 5 minutes]
3. In a lake there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the lake, how long would it take for the patch to cover half the lake? _____ days [Answer: 47 days]

Studies 2 used a modified version of these questions along with four additional questions obtained from Thomson and Oppenheimer, 2016.

1. A phone and a charger cost $110 in total. The phone costs $100 more than the charger. How much does the charger cost? _____ dollars [Answer: $5]
2. If it takes 5 workers 5 minutes to make 5 toys, how long would it take 100 workers to make 100 toys? _____ mins [Answer: 5 minutes]
3. In a field there is a patch of dandelions. Every day, the patch doubles in size. If it takes 52 days for the patch to cover the field, how long would it take for the patch to cover half the field? _____ days [Answer: 51 days]
4. If you're running a race and you pass the person in second place, what place are you in? Please respond numerically. [Answer: 2 place]
5. A farmer had 15 sheep and all but 8 died. How many are left? [Answer: 8]
6. Emily's father has three daughters. The first two are named April and May. What is the third daughter's name? [Answer: Emily]
7. How many cubic feet of dirt are there in a hole that is 3' deep x 3' wide x 3' long? [Answer: 1]

Figure 1D

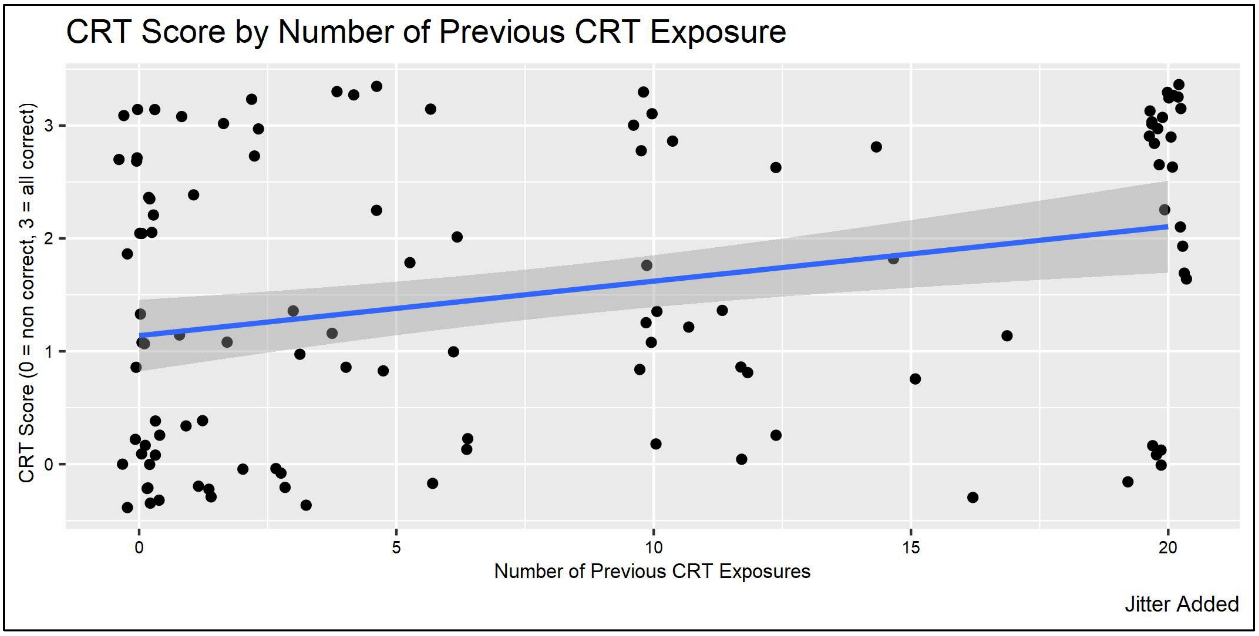
Study 1



Note. N = 113

Figure 2D

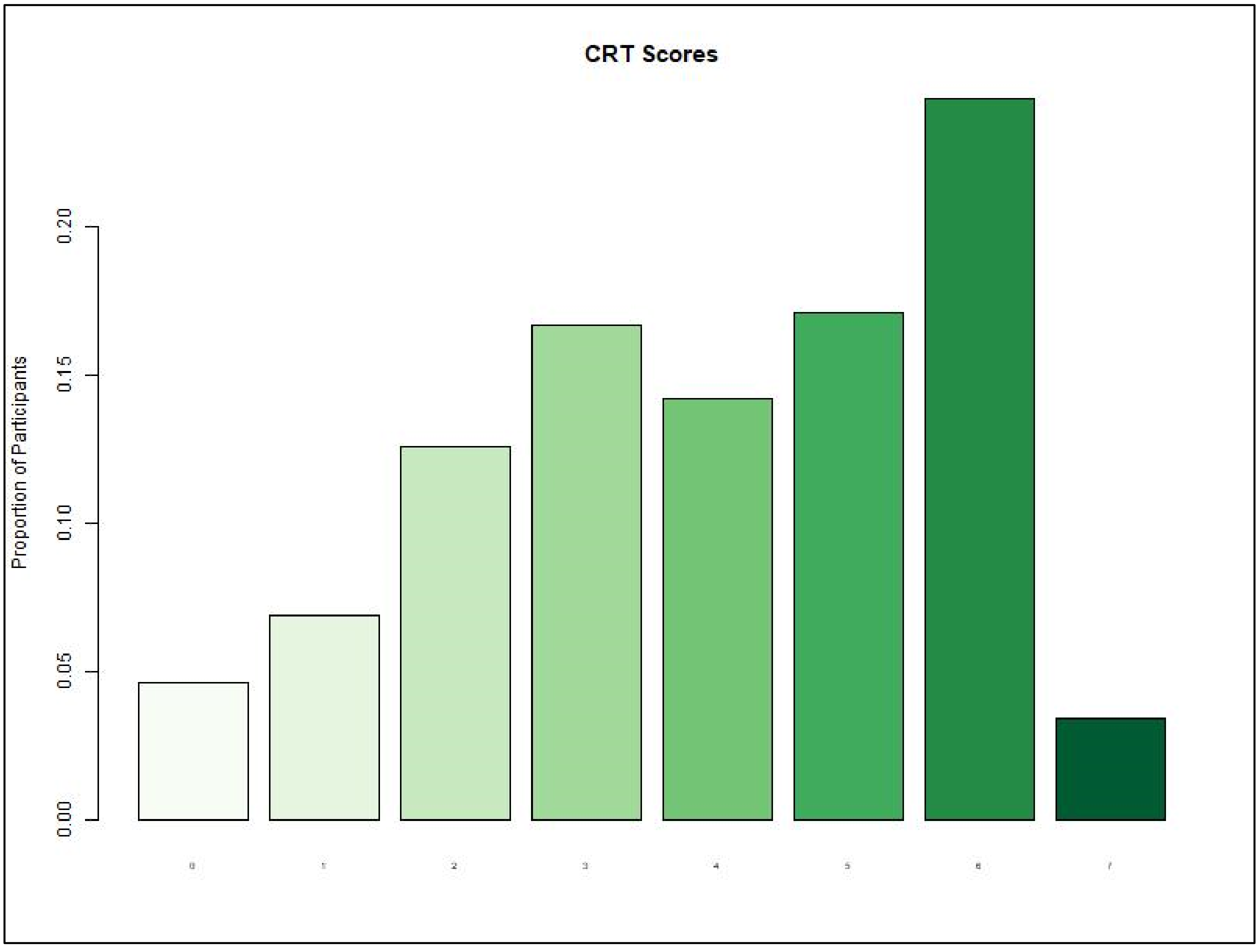
Study 1



Note. Average CRT Score = 1.14. CRT Score ~ CRT Exposure coefficient = 0.048, significant, (p = .001).

Figure 3D

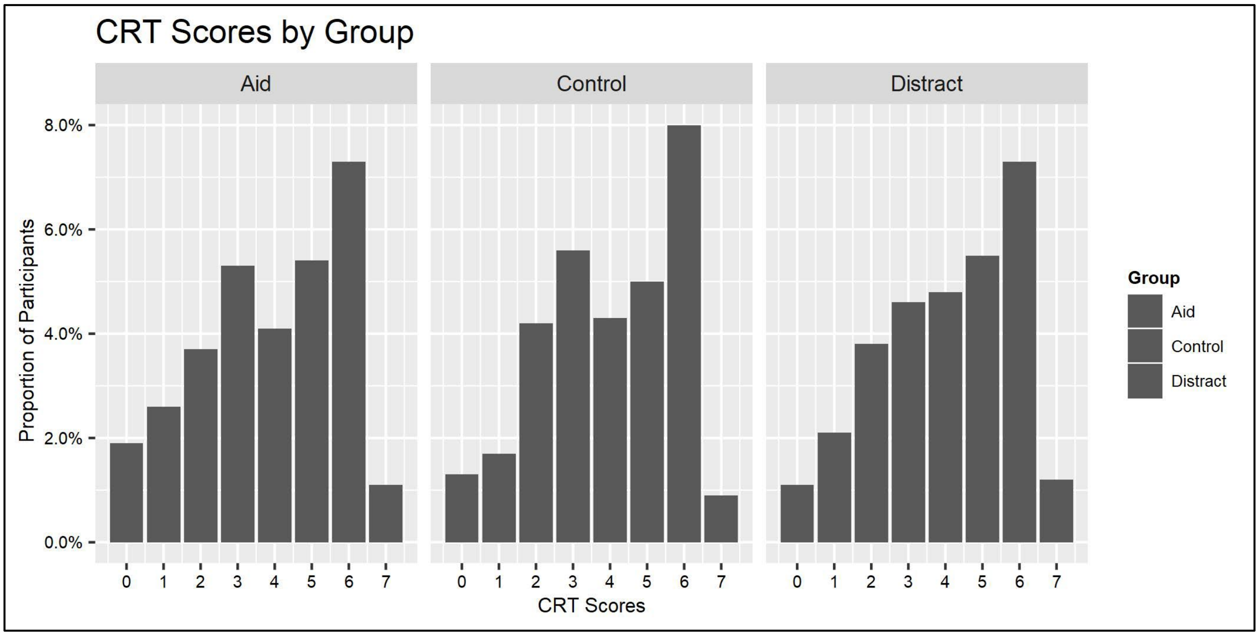
Study 2



Note. N = 924

Figure 4D

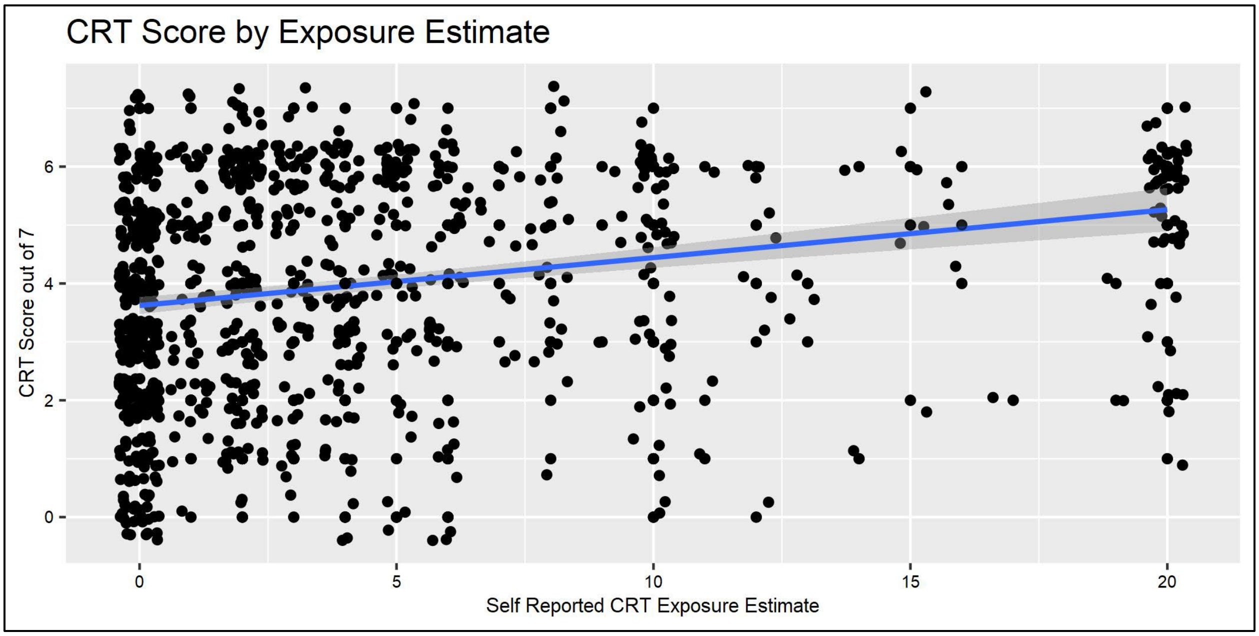
Study 2



Note. N = 924

Figure 5D

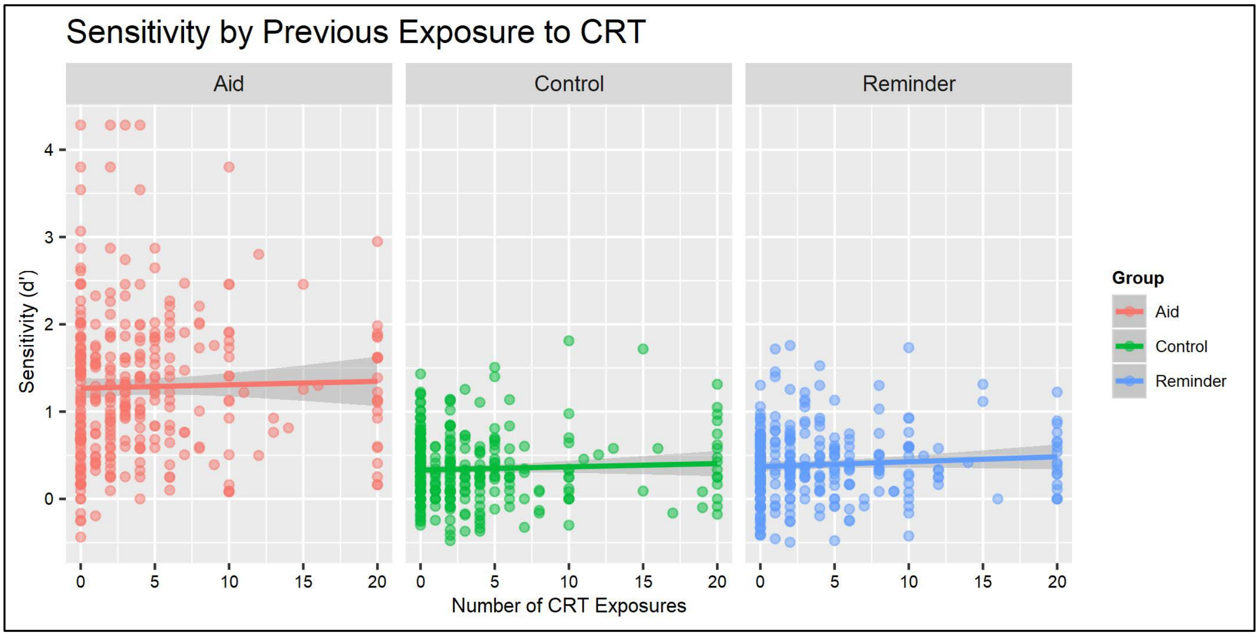
Study 2



Note. Average CRT Score = 3.62. CRT Score ~ CRT Exposure coefficient = 0.082, significant, (p < .001).

Figure 6D

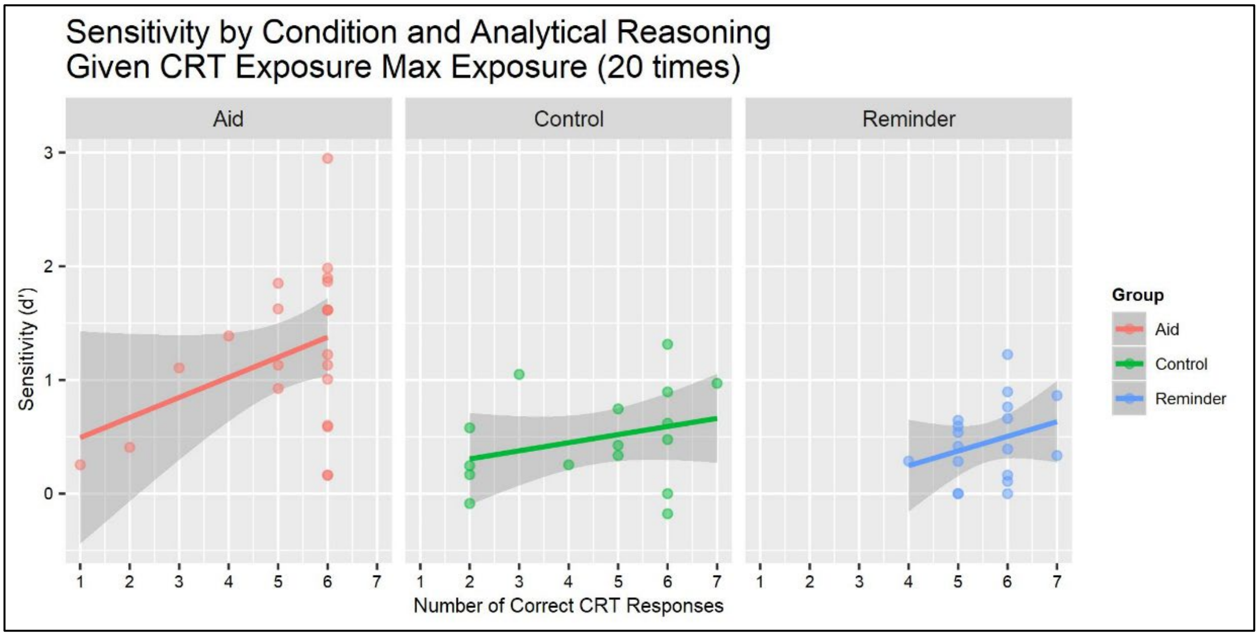
Study 2 By Group



Note. Previous Exposure ranged from 0 to 20.

Figure 7D

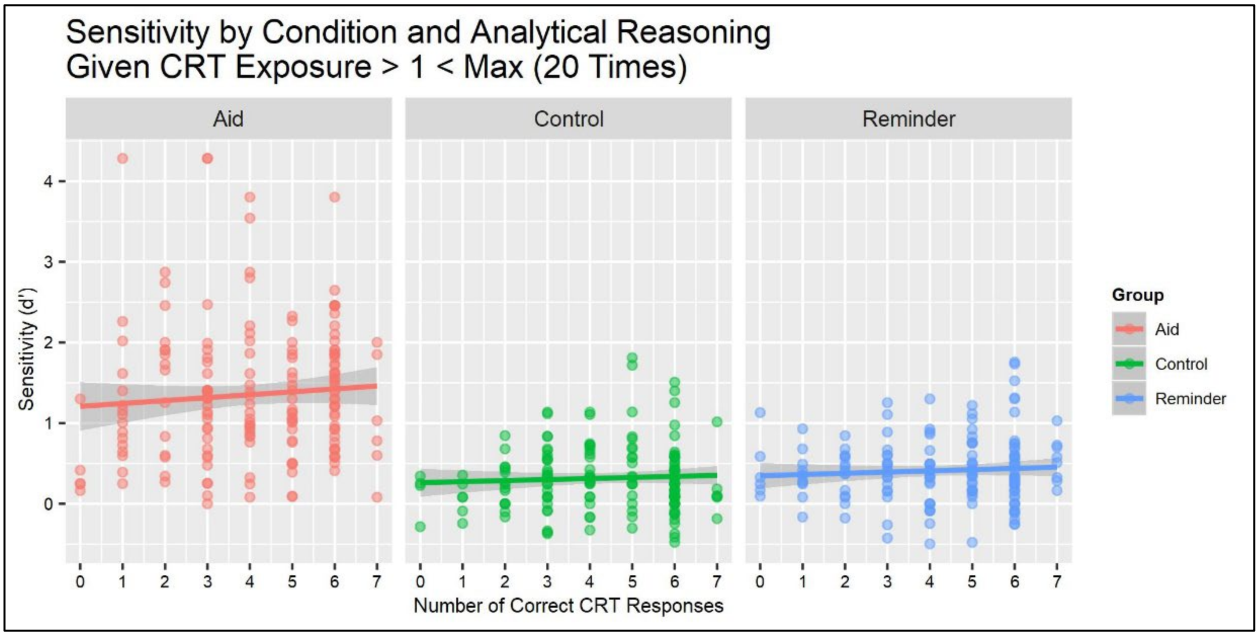
Study 2



Note. This plot represents participants who responded that had taken the CRT at least 20 times. As the max bound was 20, they may have seen it more than that.

Figure 8D

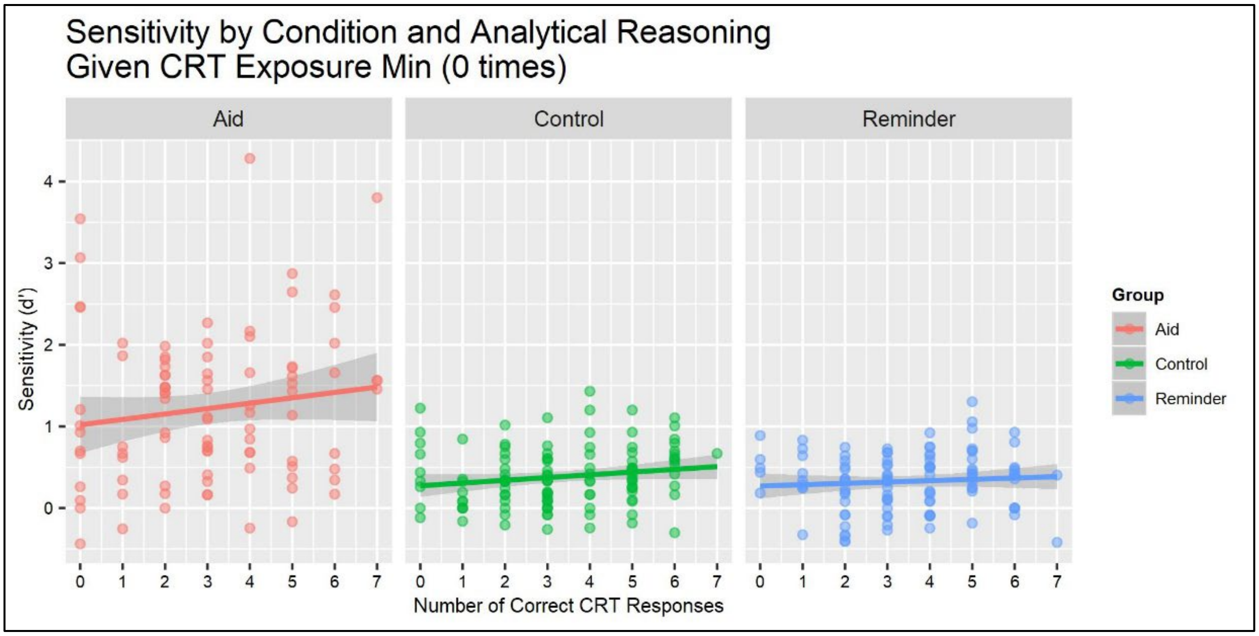
Study 2



Note. This plot represents all participants who had reported taking the CRT at least one time, but less than 20 times.

Figure 9D

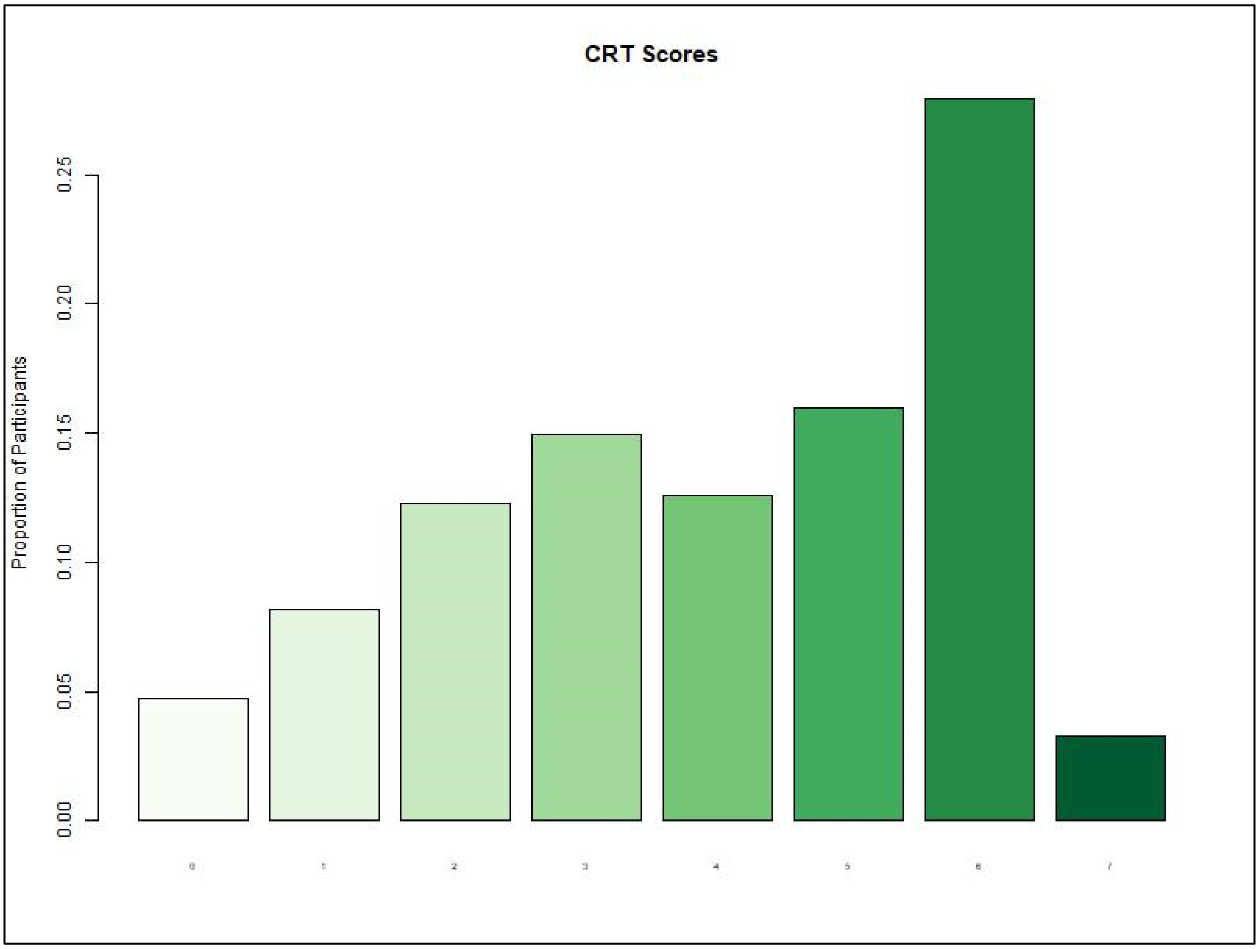
Study 2



Note. This plot represents participants who reported completing the CRT for the first time in our study.

Figure 10D

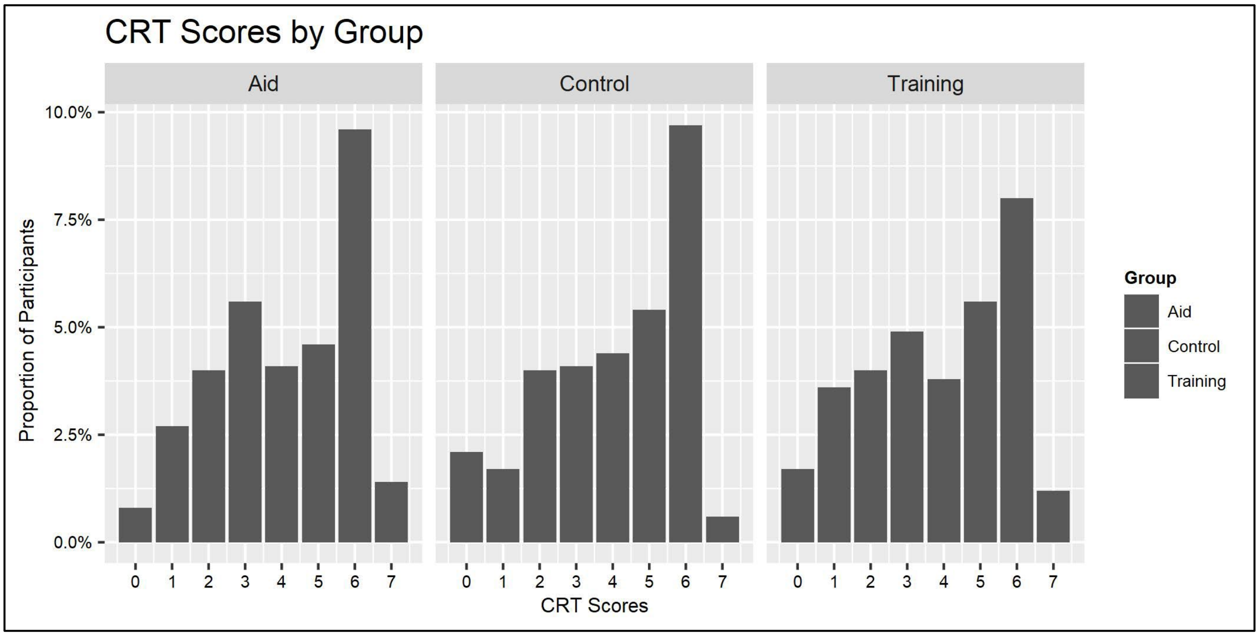
Study 3



Note. N = 976

Figure 11D

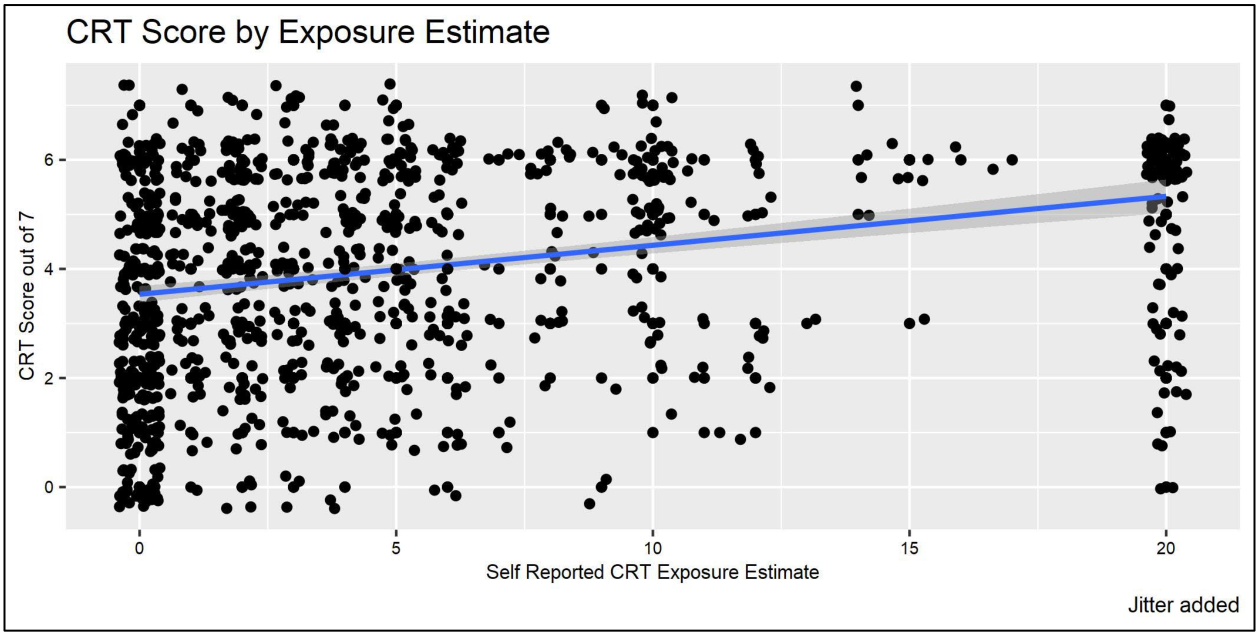
Study 3



Note. N = 976

Figure 12D

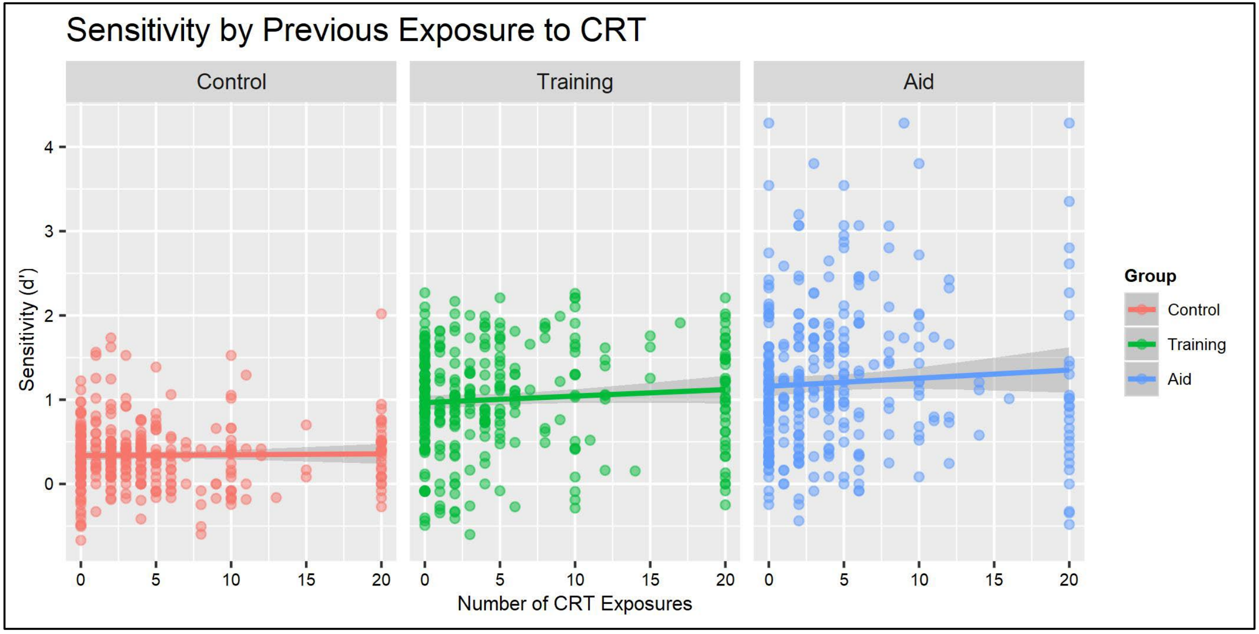
Study 3



Note. Average CRT Score = 3.54. CRT Score ~ CRT Exposure coefficient = 0.09, significant, (p < .001).

Figure 13D

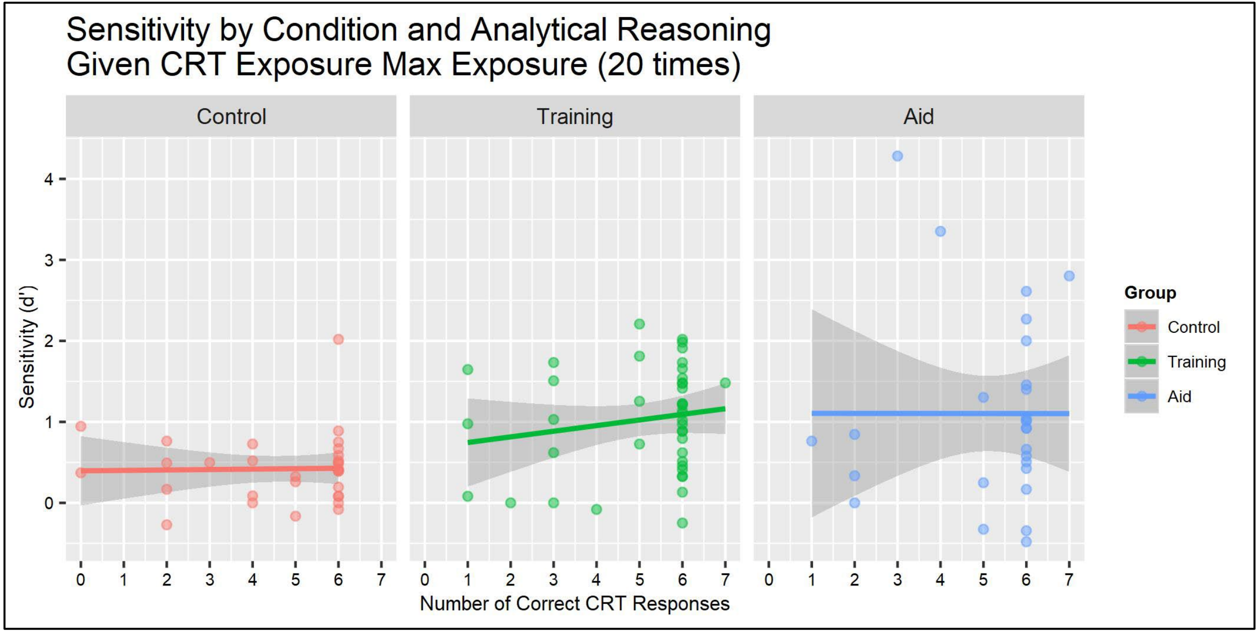
Study 3 by Group



Note. Previous Exposure ranged from 0 to 20.

Figure 14D

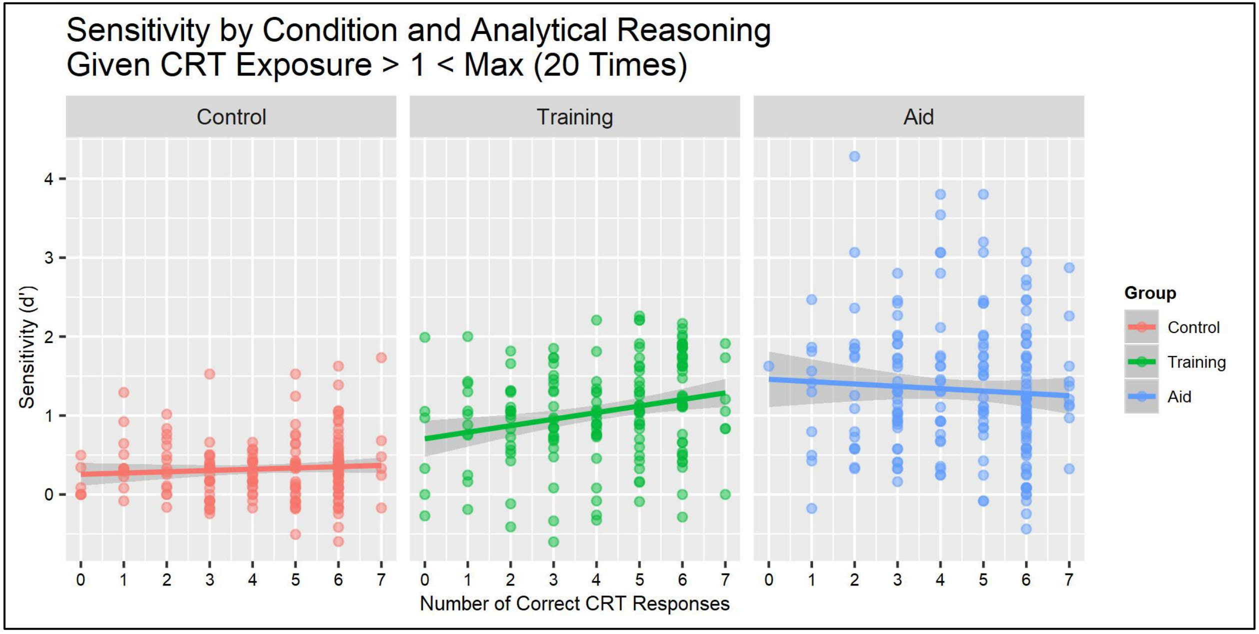
Study 3



Note. This plot represents participants who responded that had taken the CRT at least 20 times. As the max bound was 20, they may have seen it more than that.

Figure 15D

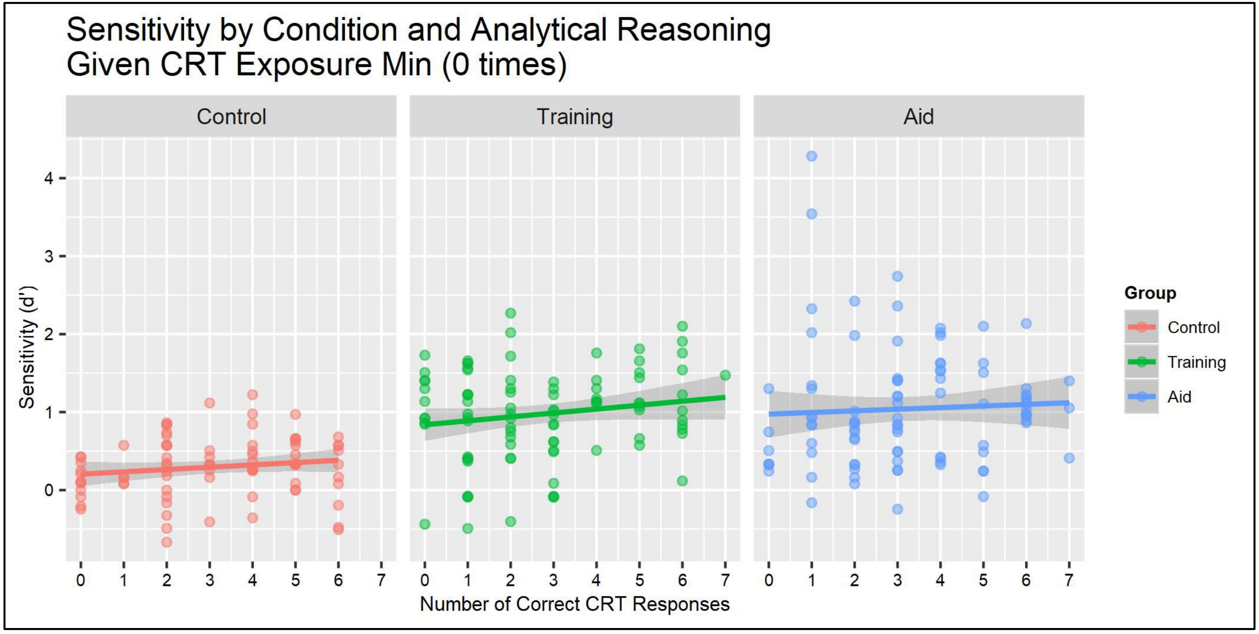
Study 3



Note. This plot represents all participants who had reported taking the CRT at least one time, but less than 20 times.

Figure 16D

Study 3



Note. This plot represents participants who reported completing the CRT for the first time in our study.

Appendix E

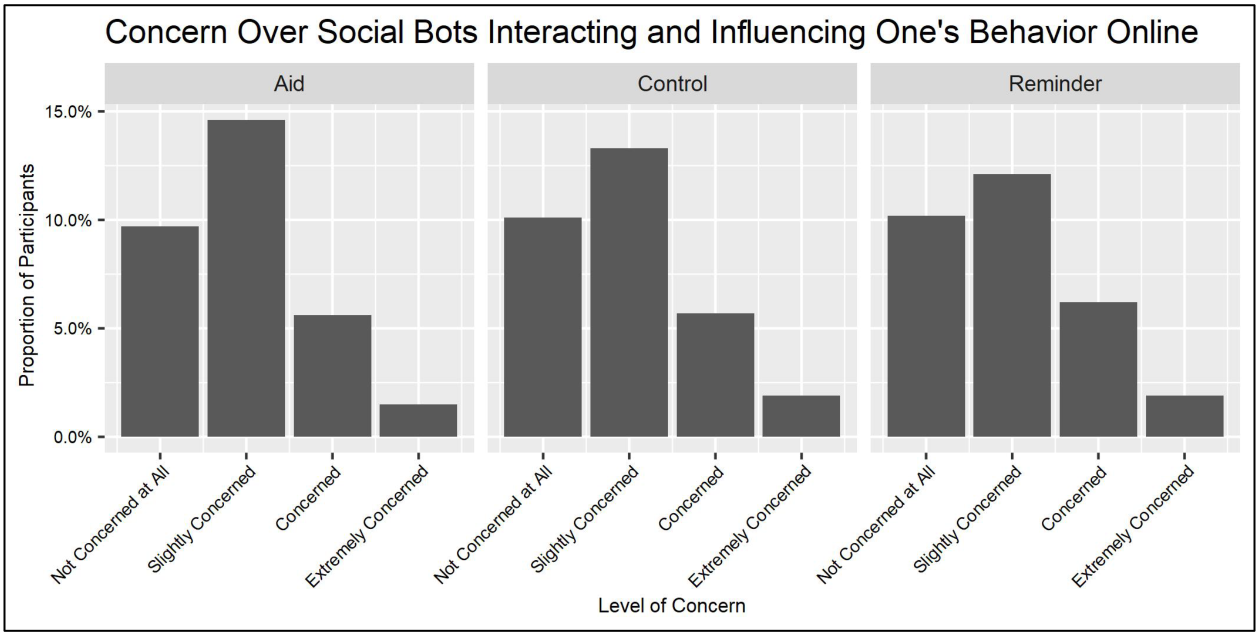
Social Bot Concern Estimates

Participants in studies 2 and 3 were asked about their concerns over social bots. Those questions and descriptive statistics of participants' responses to these questions follow.

1. Self Concern: How concerned are you about social bots affecting YOUR behavior online?
    a. Extremely concerned
    b. Concerned
    c. Slightly concerned
    d. Not concerned at all
2. Concern Others: How concerned are you about social bots affecting OTHERSs behavior online?
    a. Extremely concerned
    b. Concerned
    c. Slightly concerned
    d. Not concerned at all
3. When, in the future, do you think that social bots will cause so much harm to society that they should be controlled or banned?
    a. I do not believe social bots will ever cause society too much harm
    b. They are already causing too much harm to society
    c. 1-5 years from now
    d. 6-10 years from now
    e. 11-15 years from now
    f. 16 to 20 years from now
    g. Beyond 20 years from now

Figure 1E

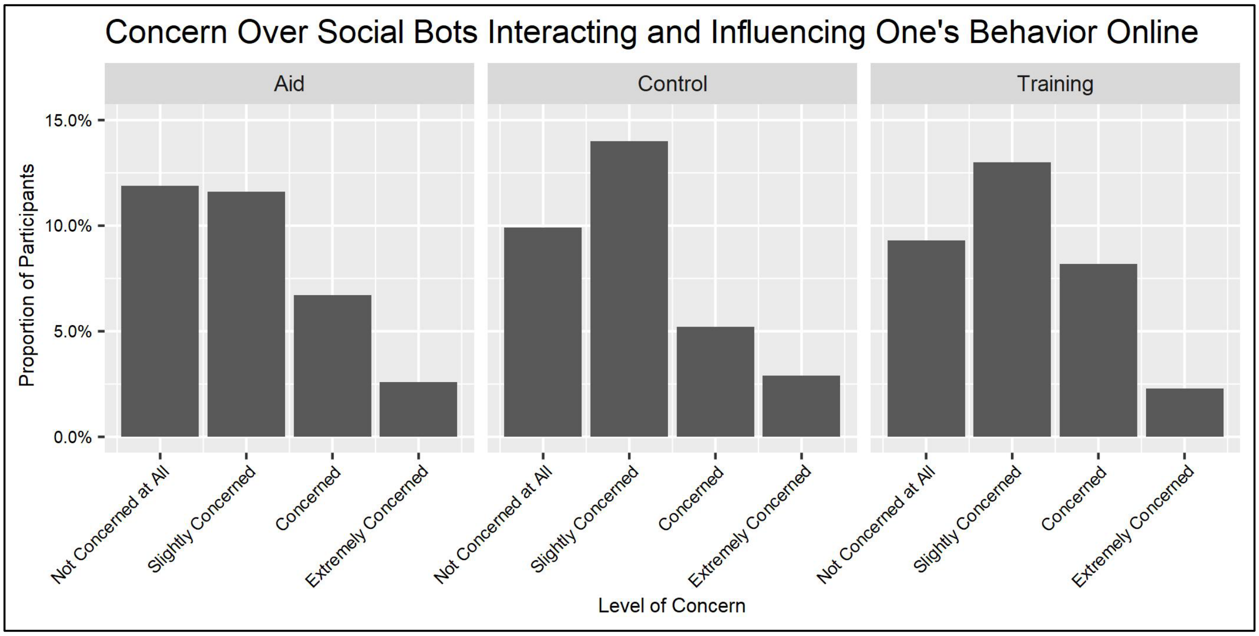
Study 2



Note. N = 924

Figure 2E

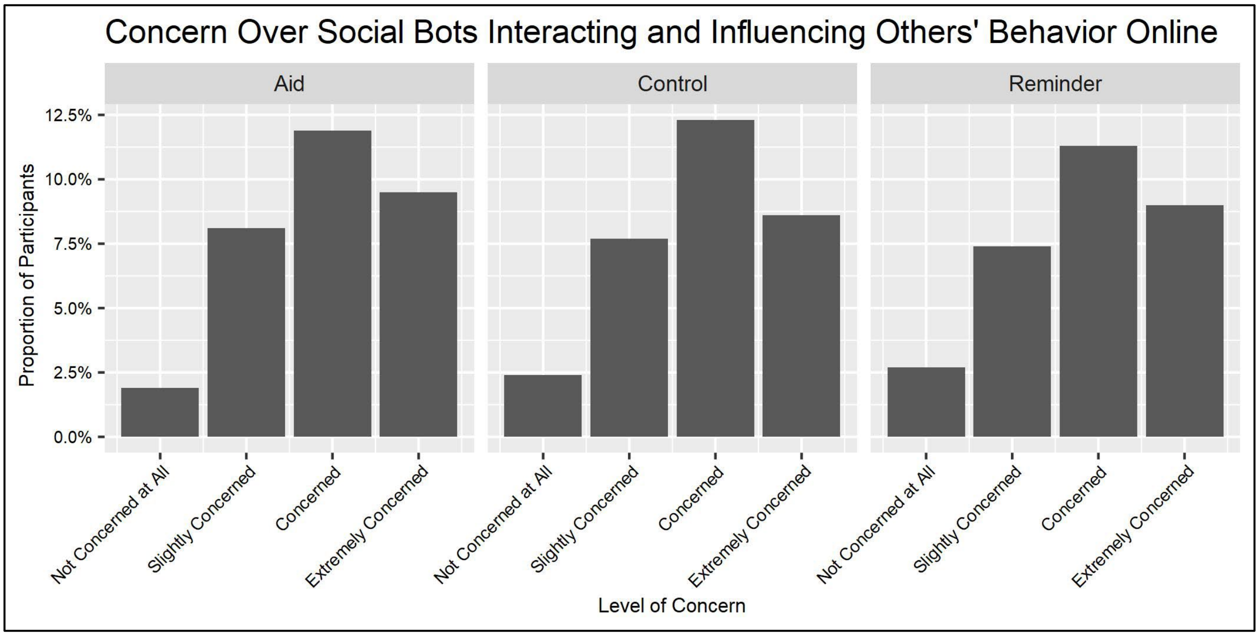
Study 3



Note. N = 976

Figure 3E

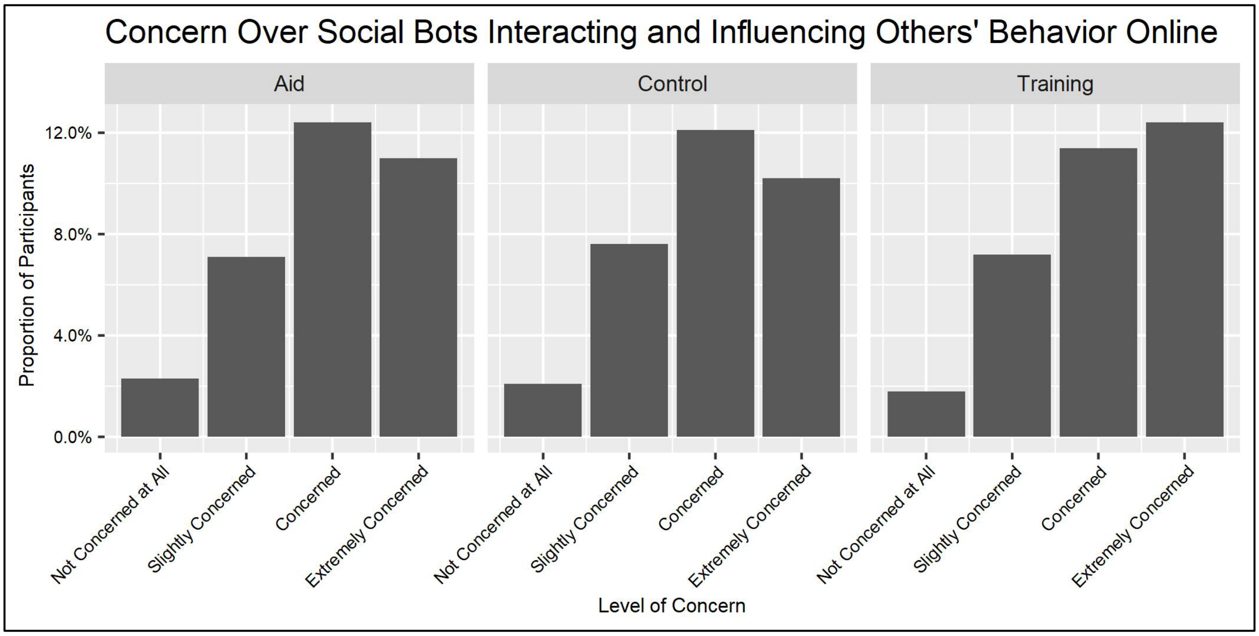
Study 2



Note. N = 924

Figure 4E

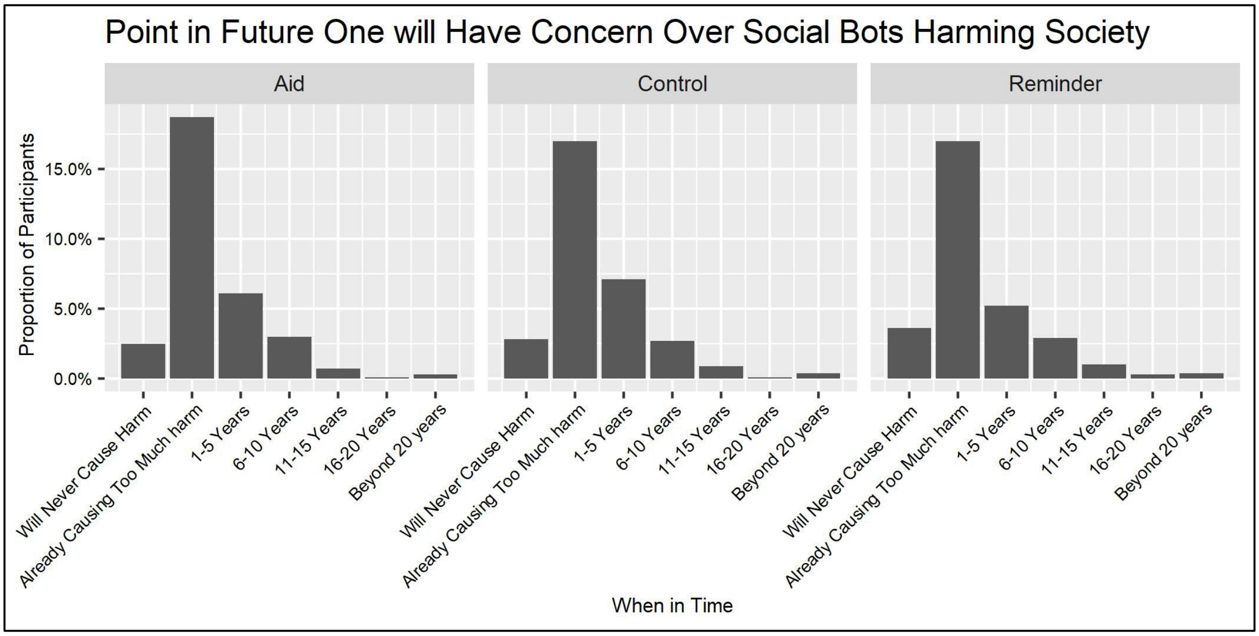
Study 3



Note. N = 976

Figure 5E

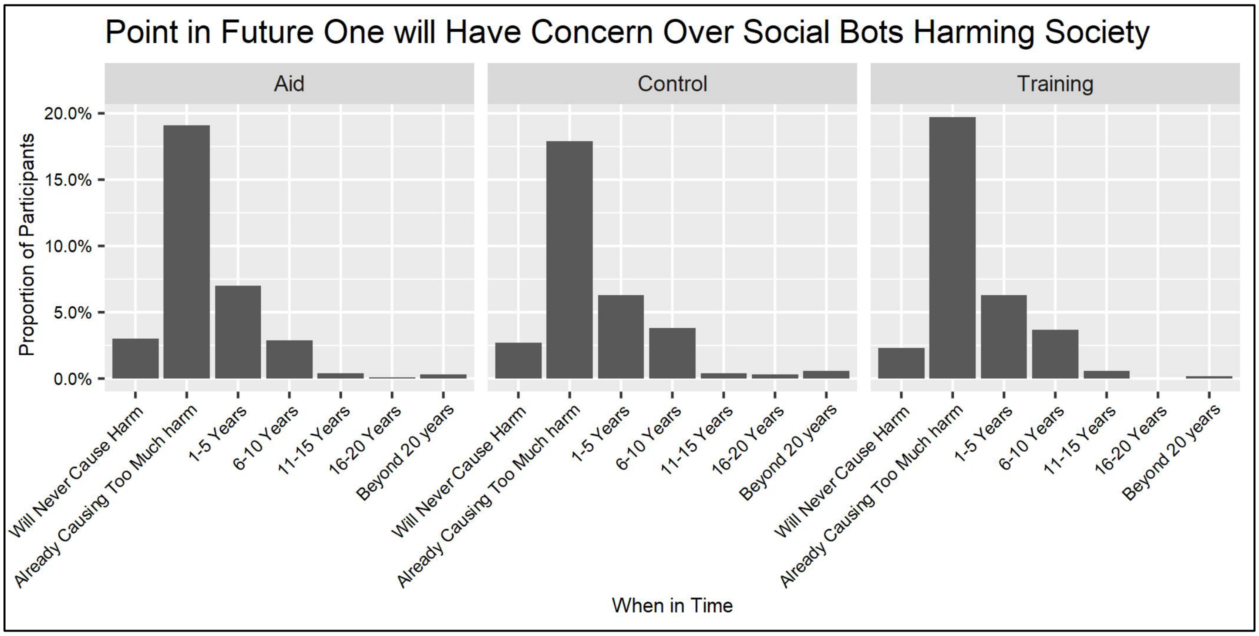
Study 2



N = 924

Figure 6E

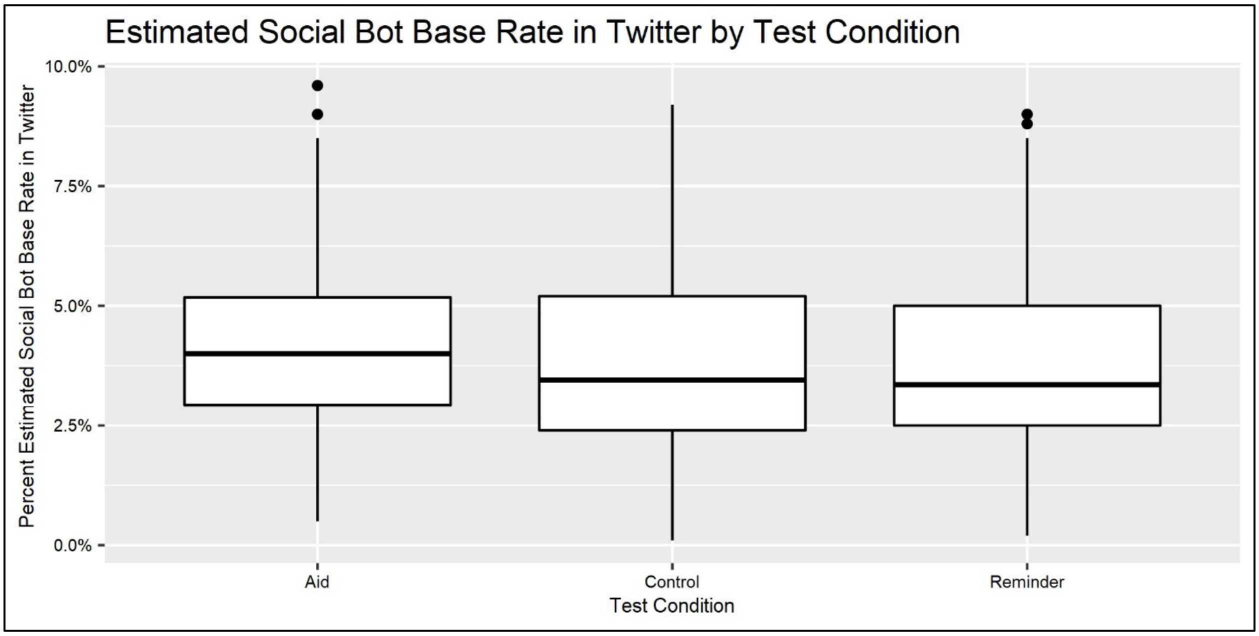
Study 3



Note. N = 976

Figure 7E

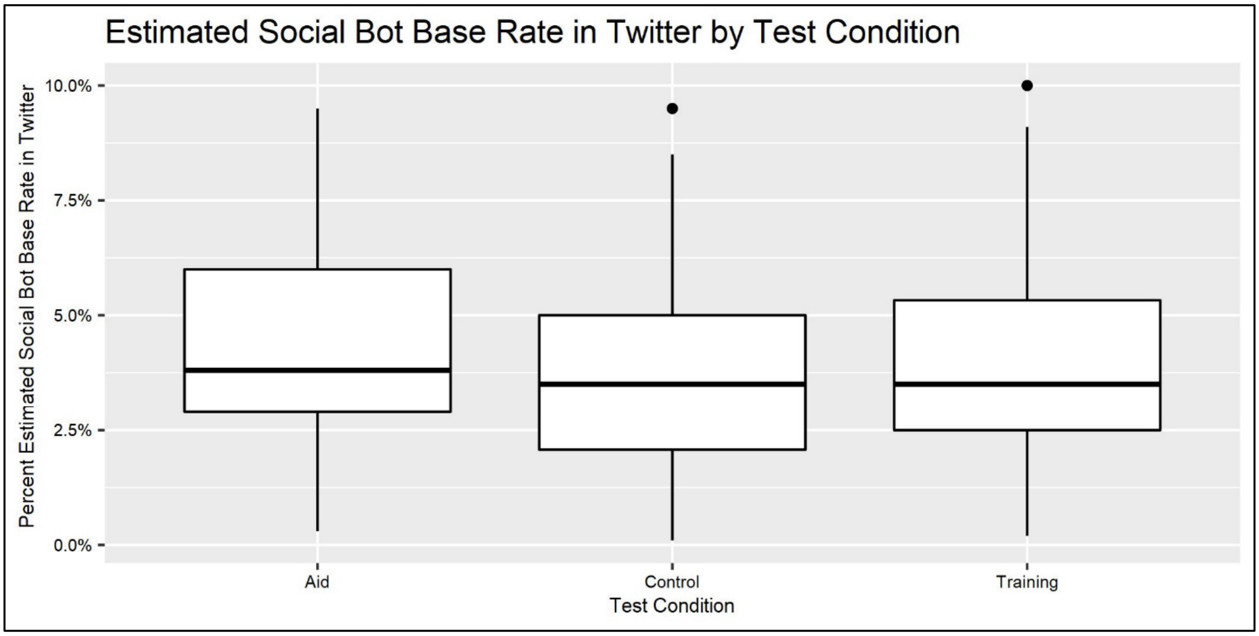
Study 2



Note. N = 928

Figure 8E

Study 3

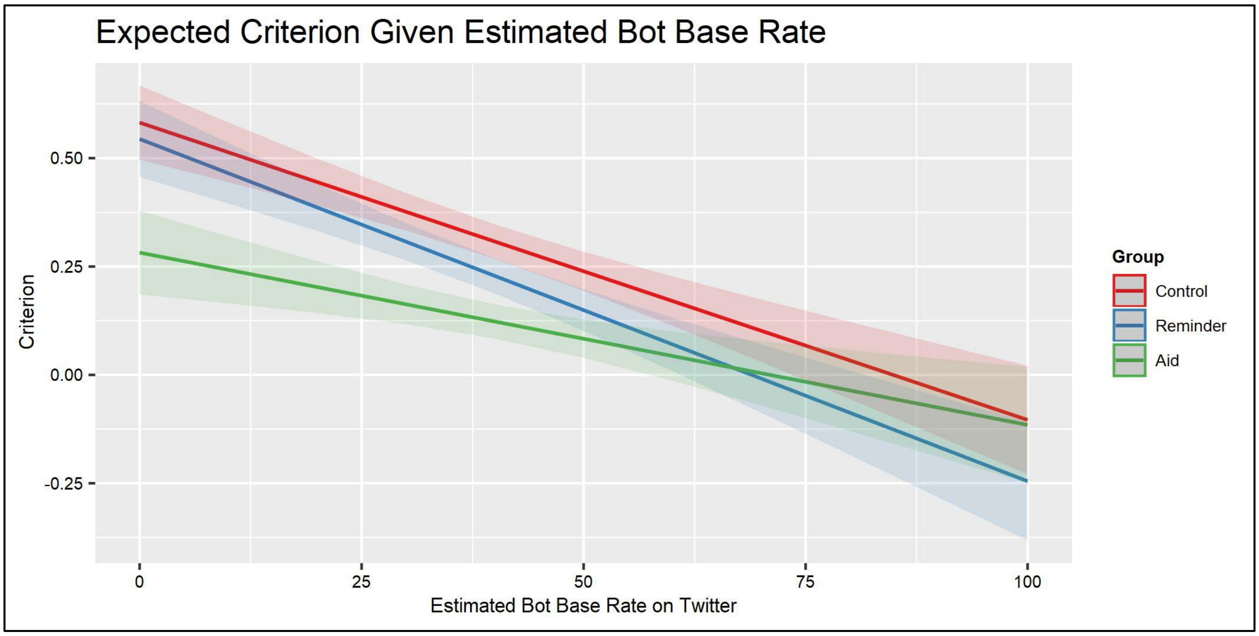


Note. N = 976

For exploratory purposes, and not included in any of study 2's or study 3's models, we examined the relationship of participants' responses to social bot estimates with their average criterions. In general, the higher one's social bot estimates were, the lower there criterion for responding 'bot' was.

Figure 9E

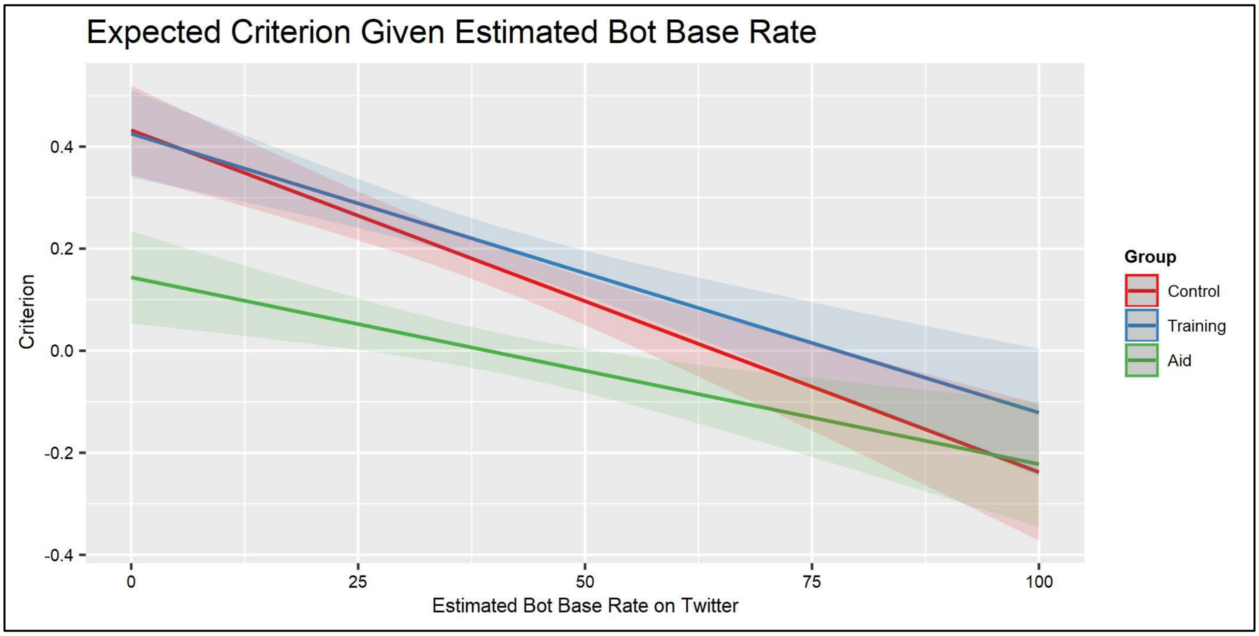
Study 2



Note. Participants were asked to estimate the percentage of personas's on Twitter that were bots. This figure plots the relationship between those estimates and a given participants' average criterion, by test condition.

Figure 10E

Study 3



Note. Participants were asked to estimate the percentage of personas's on Twitter that were bots. This figure plots the relationship between those estimates and a given participants' average criterion, by test condition.

Appendix F

Willingness to Pay

In studies 2 and 3, participants were asked the following willingness to pay and willingness to delay questions. Those questions follow.

1. What is the maximum amount you would be willing to pay monthly for a social bot detection tool that told you which online personas are social bots? [Responses were not bound]
2. If the social bot detection system were free, but caused a delay in loading Twitter persona profiles, what is the maximum delay that you would find acceptable?
   a. 1 second
   b. 2-5 seconds
   c. 5-10 seconds
   d. 10-30 seconds
   e. 30 seconds or more

Figure 1F

Study 2 Social Media Experience – Exploratory WTP Relationships

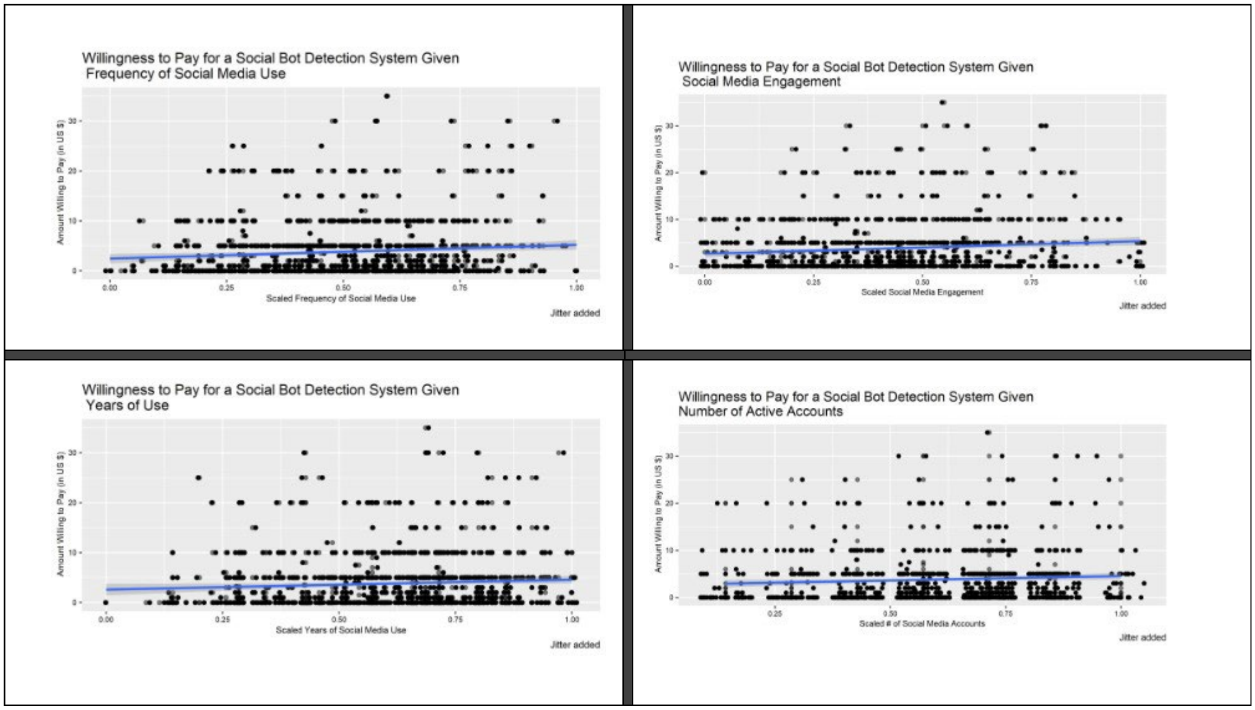
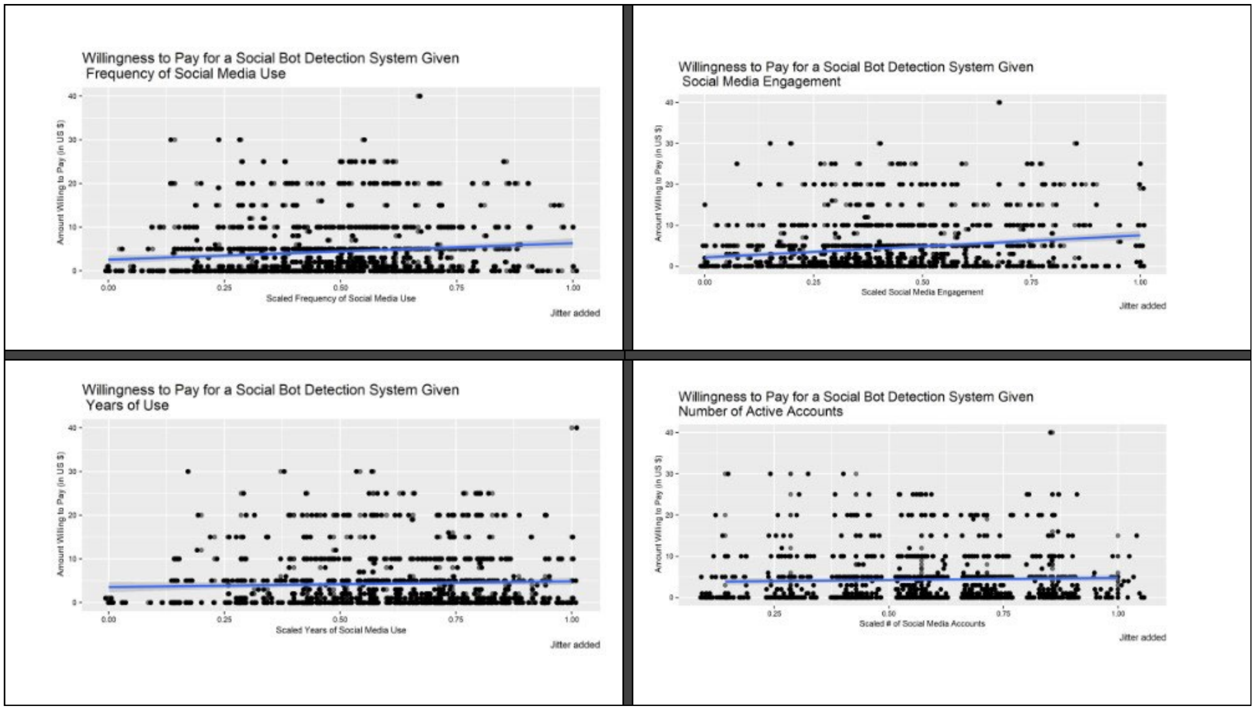


Figure 2F

Study 3 Social Media Experience – Exploratory WTP Relationships



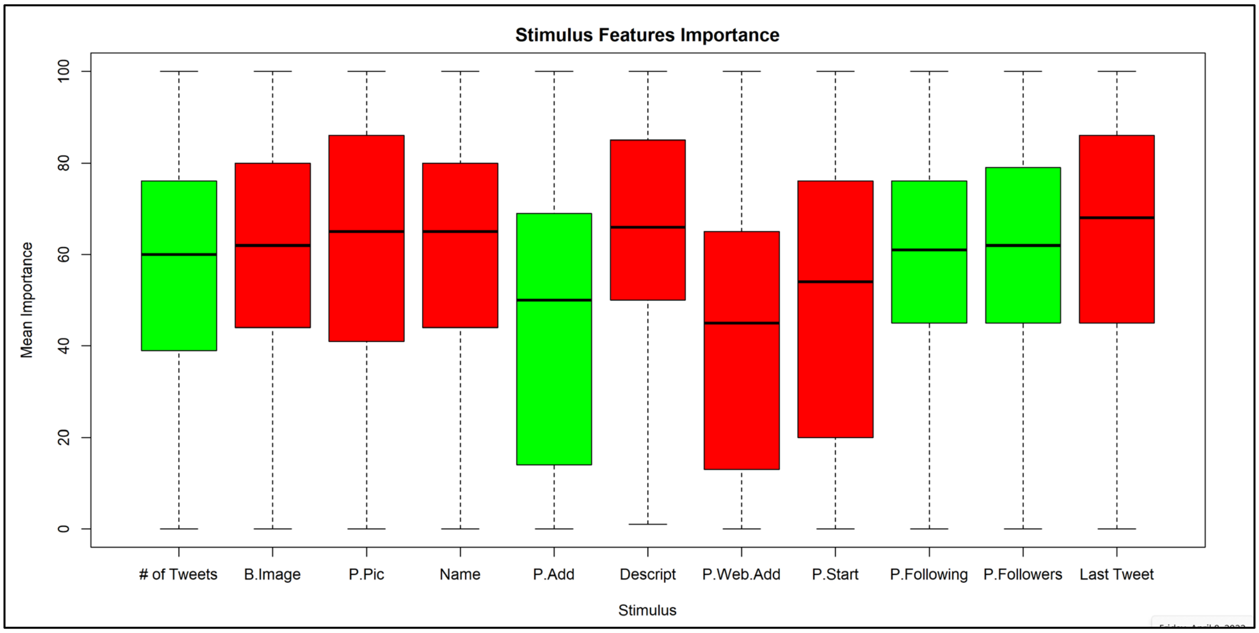
Appendix G

Stimulus Feature Ratings

Participants in each of the studies were shown a Twitter profile with features numbered and asked, “how important each of the following features was in making your determination whether you believed a persona was a human or a bot.” Using a scale from 0, “Very Unimportant” to 100, “Very Important” participants responded. The following figures display those results and relationships between participants’ ratings of the importance of these features and their sensitivity across the SDT trials for features of interest.

Figure 1G

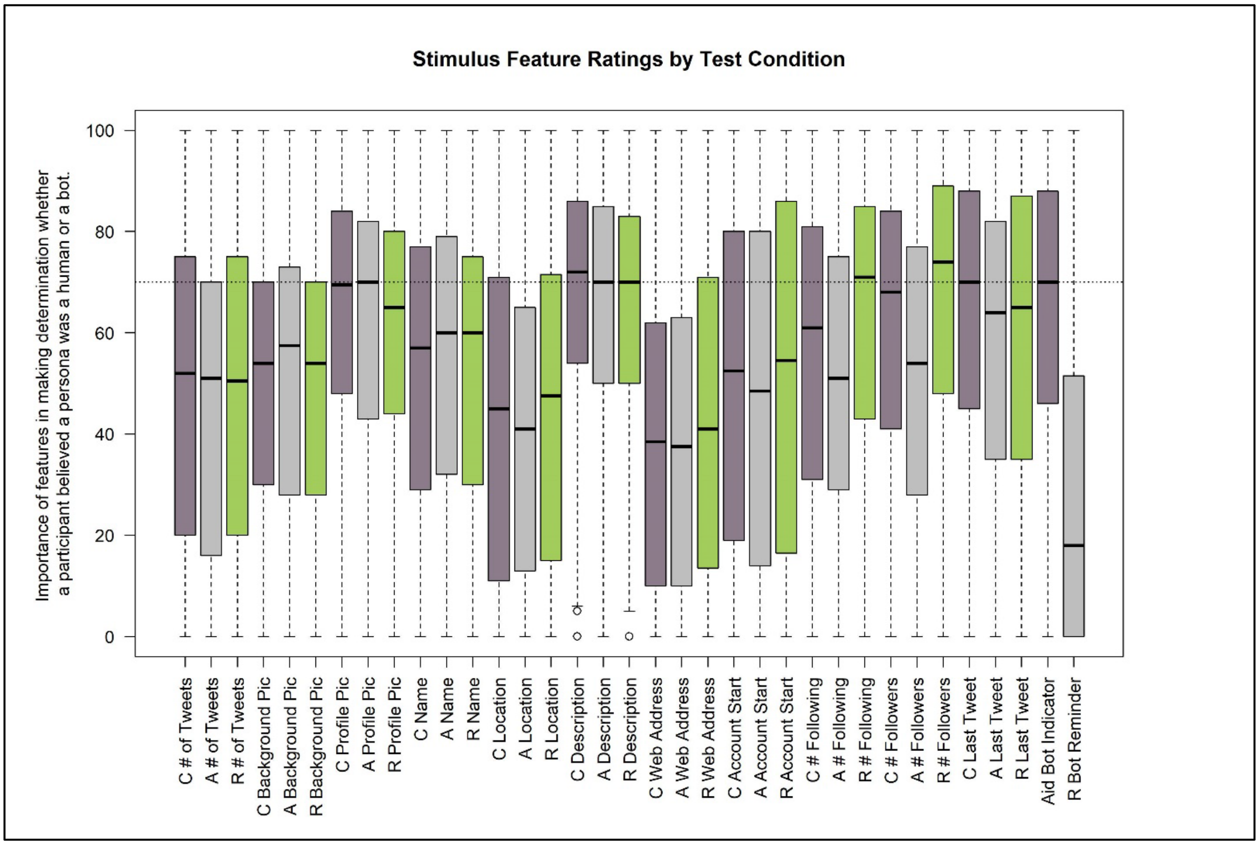
Study 1



Note. The features in red are those that can be controlled by account managers. The features in green are those tracked by Twitter and displayed on a profile automatically.

Figure 2G

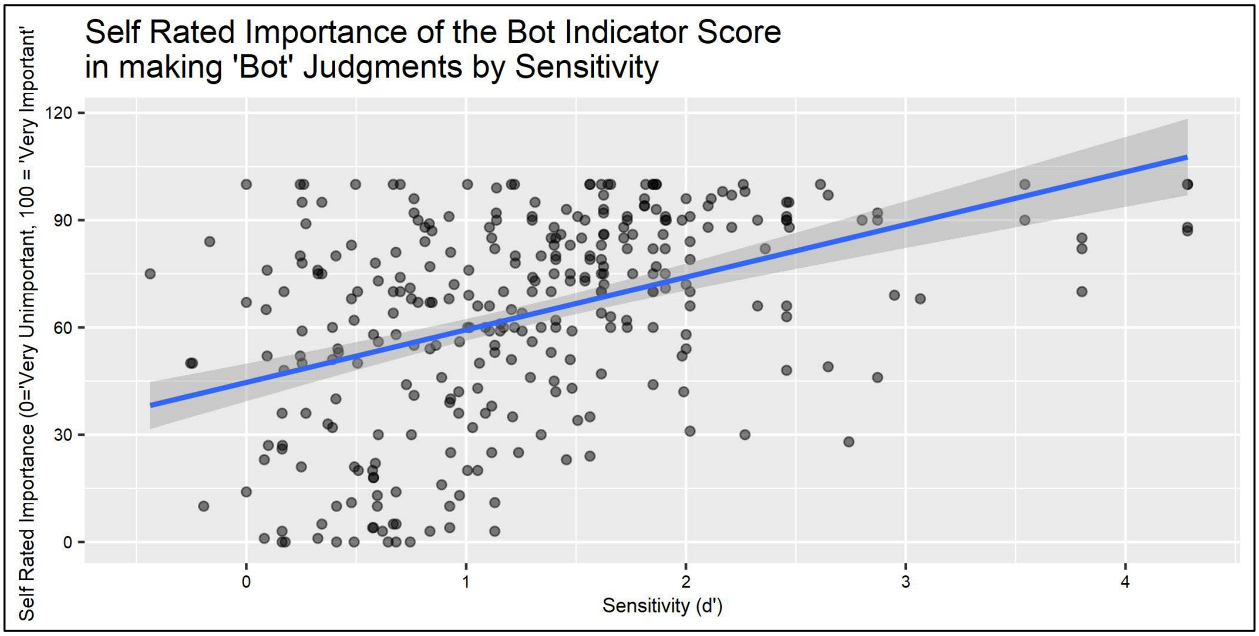
Study 2



Note. ‘C’ stands for control group, ‘A’ stands for aid group, and ‘R’ stands for reminder group. The horizontal line is equal to the average rating for the bot indicator score feature.

Figure 3G

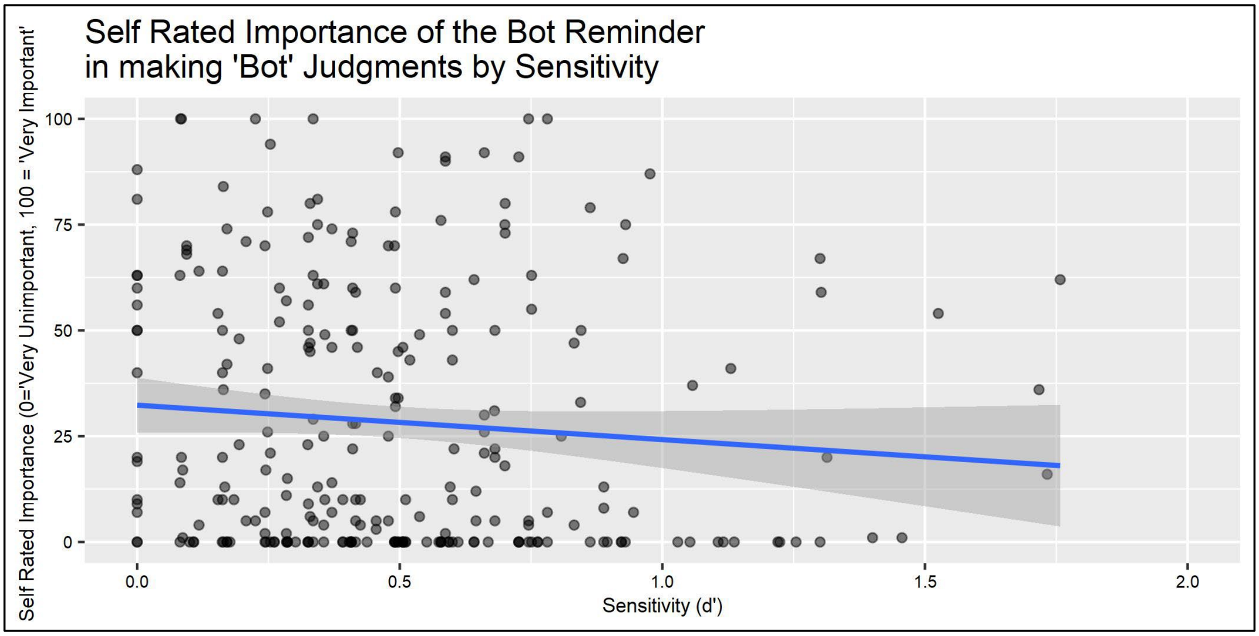
Study 2



Note. Participants in the aid test condition, who rated the bot indicator score important

Figure 4G

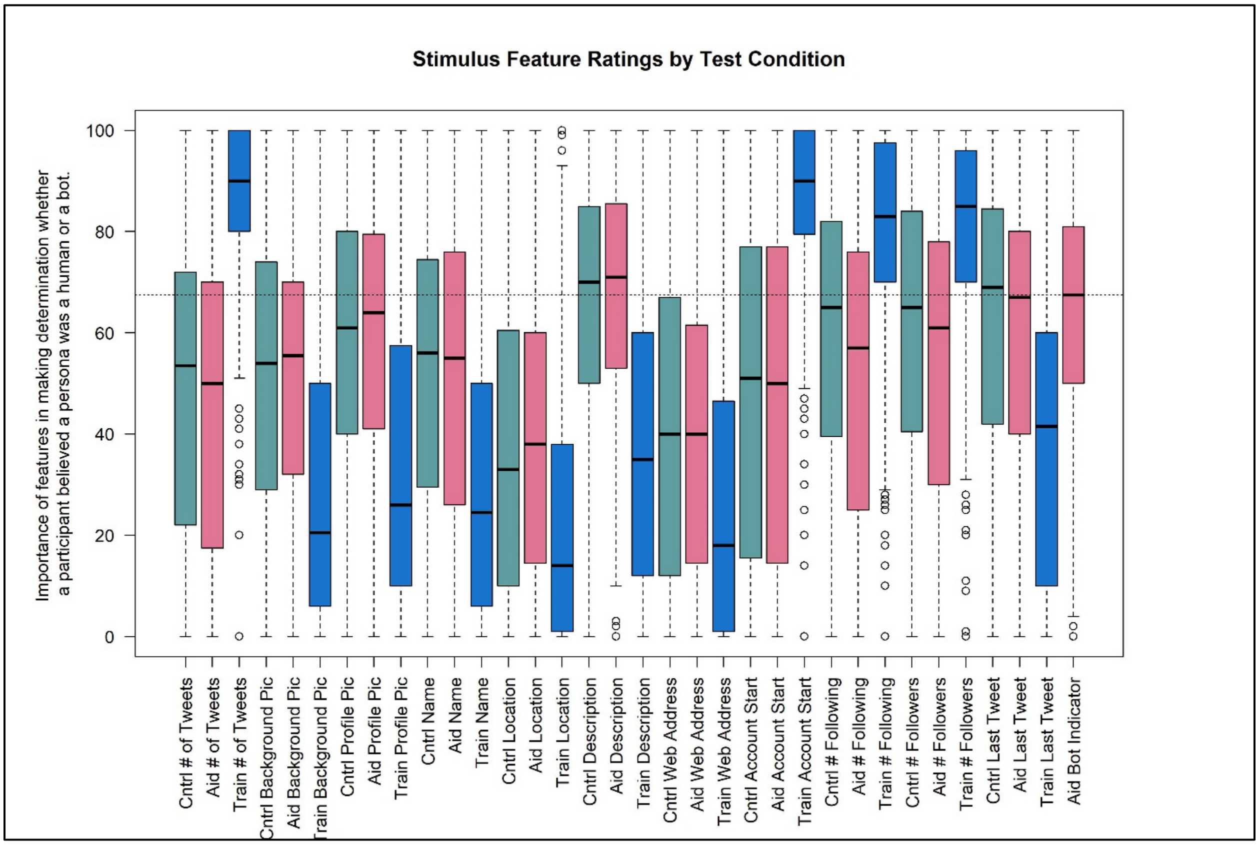
Study 2



Note. Participants in the reminder condition were shown the prompt “Look for bo cues” near a persona’s profile name.

Figure 5G

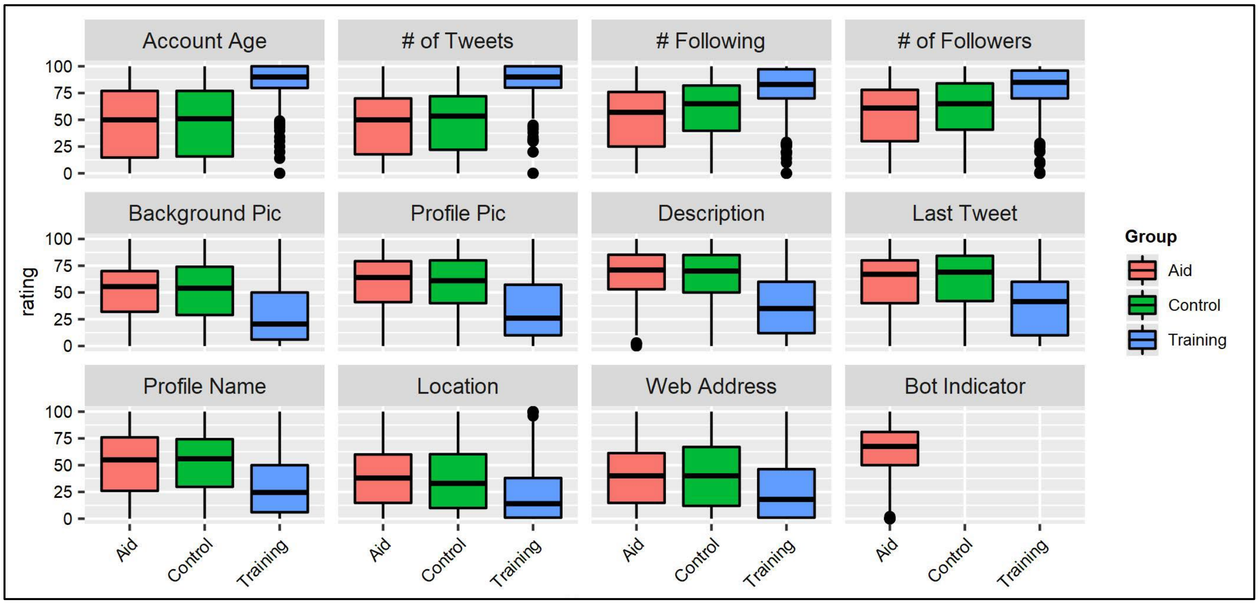
Study 3



Note. The horizontal line is equal to the average rating for the bot indicator score feature. The four features emphasized in the training protocol provided to the training test group were, # of Tweets, Account Start, # Followers, and # Following.

Figure 6G

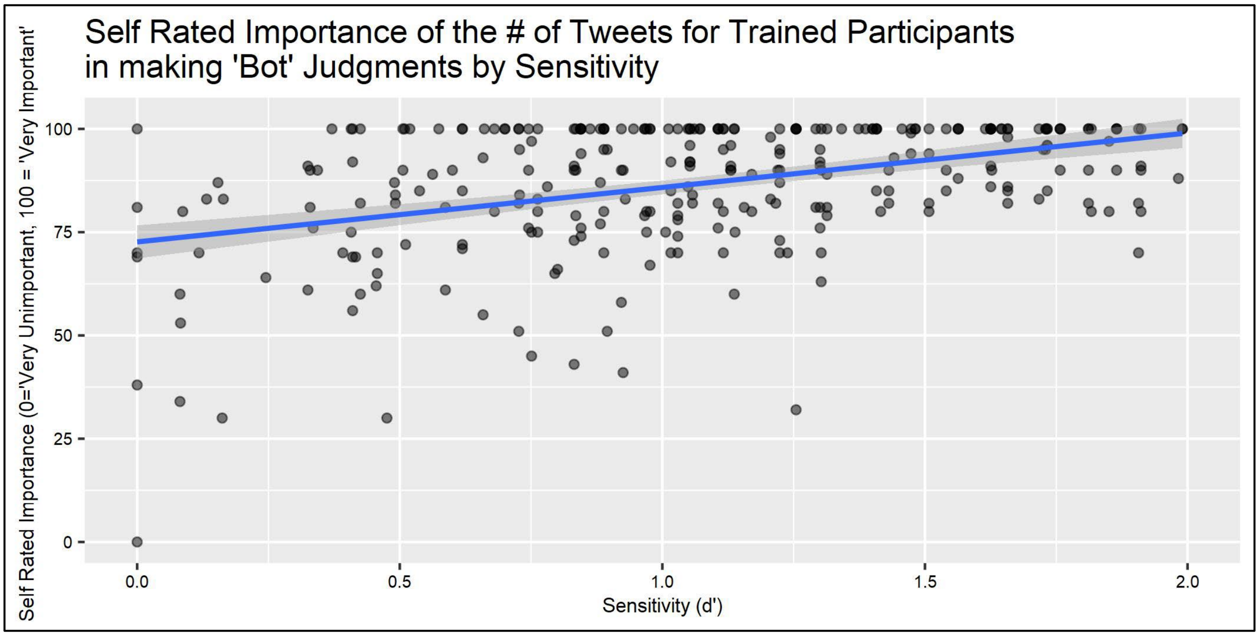
Study 3 Stimulus Feature Ratings by Test Condition Grouped



Note. This figure shows the same results as Figure 5 arranged by profile feature rather than by test condition.

Figure 7G

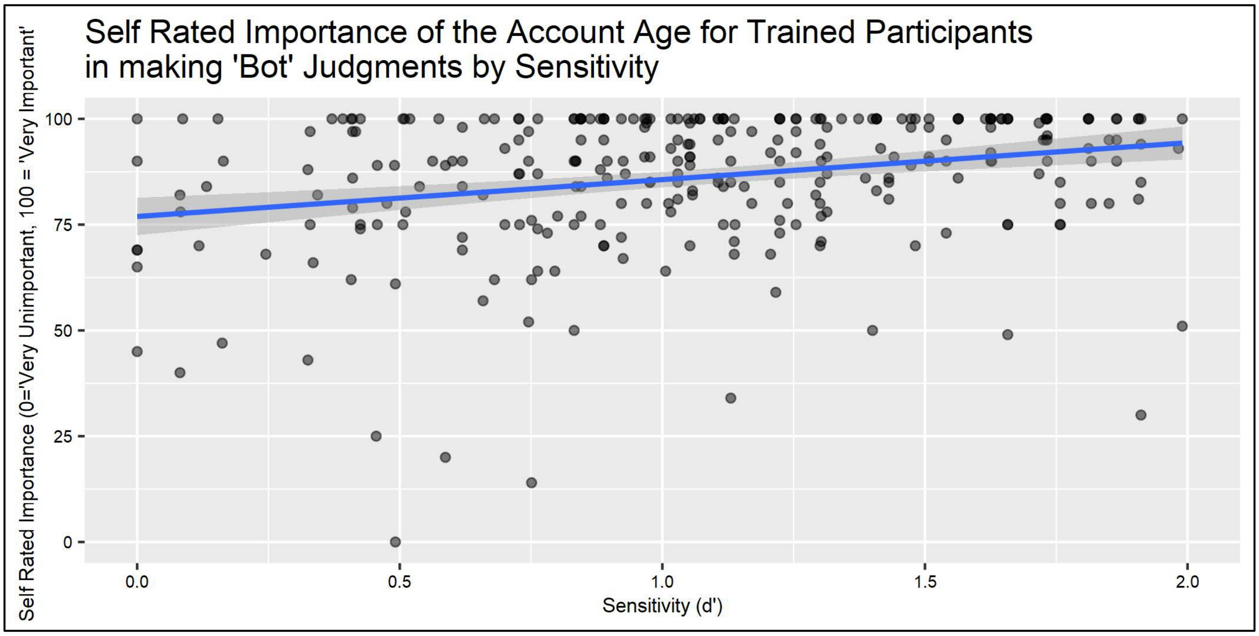
Study 3



Note. The positive correlation between a participant's importance rating for the feature # of Tweets and a participant's sensitivity was found to be significant (p < 0.001) in a linear model that included # of Tweets, account age, # Following, and # of Followers as predictor variables.

Figure 8G

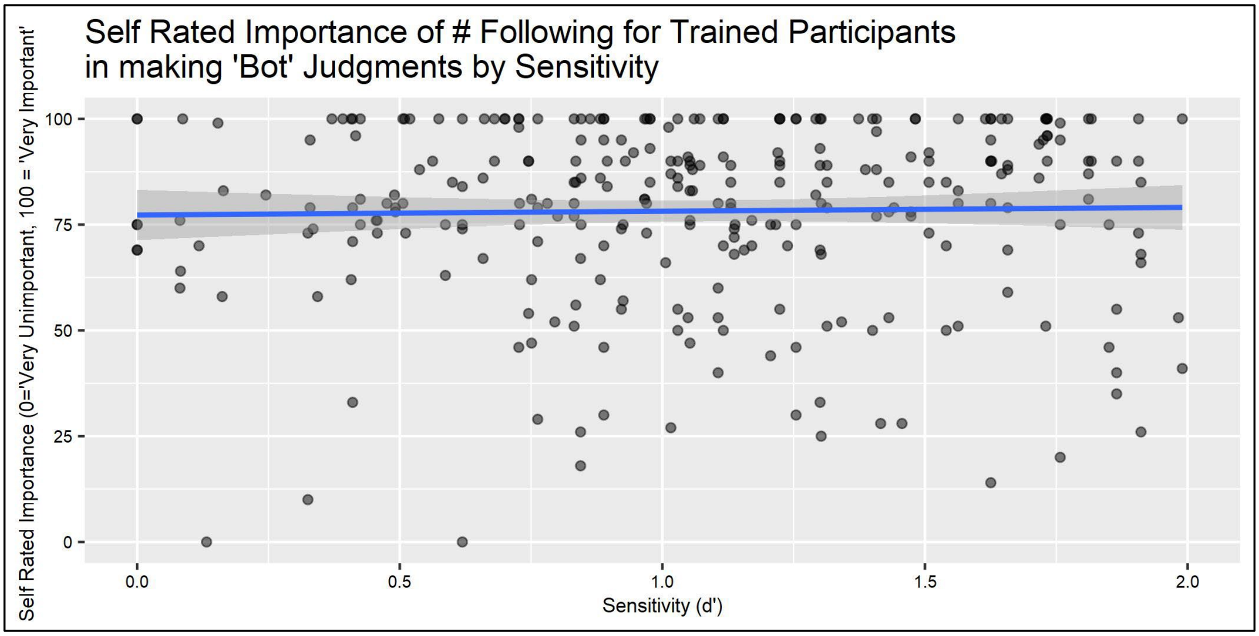
Study 3



Note. The positive correlation between a participant's importance rating for the feature Account Age and a participant's sensitivity approached significance (p = 0.095) in a linear model that included # of Tweets, account age, # Following, and # of Followers as predictor variables.

Figure 9G

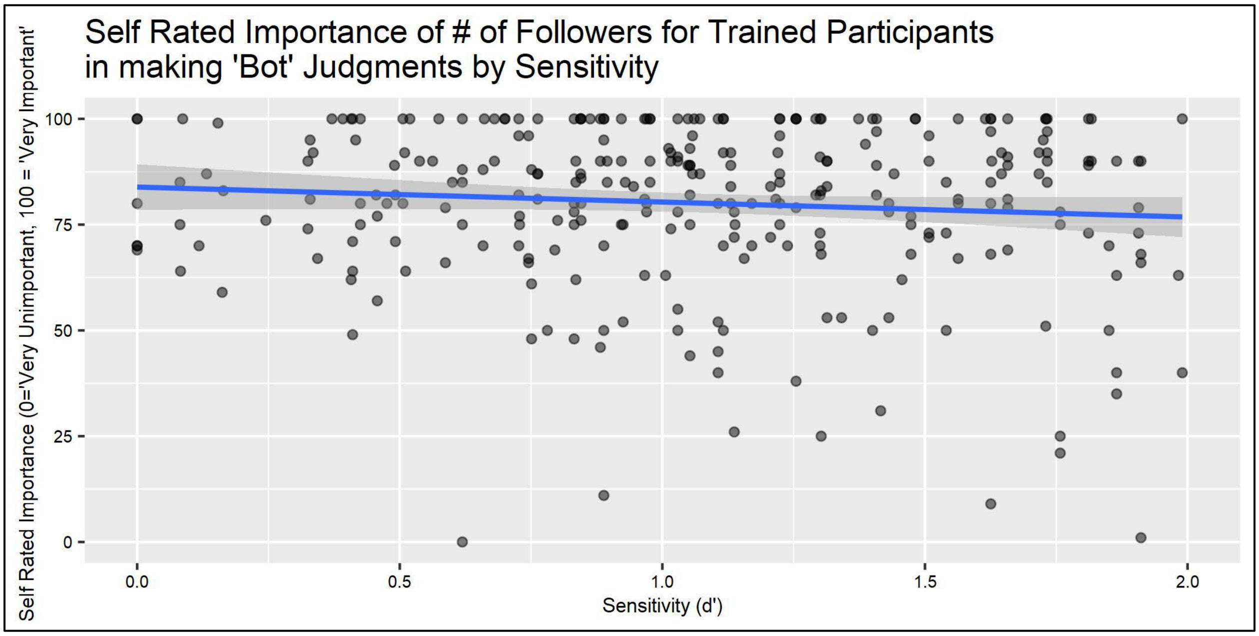
Study 3



Note. The correlation between a participant's importance rating for the feature # Following and a participant's sensitivity was not significant in a linear model that included # of Tweets, account age, # Following, and # of Followers as predictor variables.

Figure 10G

Study 3



Note. The negative correlation between a participant's importance rating for the feature # of Followers and a participant's sensitivity was significant (p < 0.001) in a linear model that included # of Tweets, account age, # Following, and # of Followers as predictor variables.

Appendix H

Training Protocol Normative Machine Learning Estimates

Existing social bot detection algorithms use a dozen or so predictive features of individual users and complex network structure data. Average users would likely struggle to make meaningful sense of this information, even if it was readily available. Furthermore, to perform a thorough analysis of the network structures of social bots, sophisticated tools are required.

Our previous findings demonstrated that participants had sensitivity to bot signals. Therefore, sufficient information must have been available (i.e., signal strength) present amongst the set of features found within the profiles to inform participants' responses. However, it is unclear which features best aided their judgments.

The stimuli used in these studies, and their associated bot indicator scores, were determined by Bot Hunter's Tier 1 Model. This model employs a random forest machine-learning algorithm to base its predictions based on user-defined features (we describe these features below). These are the same features that one could surmise by examining Twitter users' profiles.

We conducted two machine learning experiments using Bot Hunter's Tier 1 Model predictions to determine if a sufficient bot signal resided within a subset of Twitter-managed features. The first experiment trained four machine learning algorithms using platform-managed features that would indicate high amplification of information and evaluated their overall accuracy.

Social bots are employed principally to amplify narratives and influence information operations. Therefore, we conjectured that the total number of Tweets an account produced over its lifetime, combined with the age of the account, would produce a reasonable estimate of high versus low social bot amplification. Additionally, social bot developers also seek large audiences for their content. Therefore, we hypothesized that the number of accounts a persona followed and followed would also indicate high versus low social network reach, which should be correlated with the likelihood a persona is a social bot.

Using these predictive attributes in Experiment 1, we tested whether there was sufficient information provided by knowing the age of an account, the total number of Tweets it had produced in that time, along with the size of its network, whereby users could have improved accuracy well above chance levels (i.e., 50% accuracy).

The second experiment used the same four machine learning algorithms as in Experiment 1, and the same training and test data, to compare the accuracy of their predictions when limited to the remaining user-managed and less relevant platform-managed features incorporated in Bot Hunter's Tier 1 Model. We then compare and discuss the performance of these algorithms relative to our objectives of determining suitable heuristics with which to train users.

Experiment 1

Data Preparation. Dr. Beskow provided the data used to train the models. We incorporated stimuli from this data set in study 1. This data set included raw Twitter user data of approximately 4,449 accounts active during the 2018 U.S. mid-term election. This data set used the Bot Hunter social bot predictions as the outcome variable in each trained model. We used Weka (Frank, Hall, & Witten, 2016) for model development. To obtain a prediction for each of the 4,449 personas, we performed 10-fold cross-validation.

Due to their differences in approach to interpretability, we chose four algorithms to test: Naïve Bayes, logistic regression, a J48 (C4.5 algorithm) Decision Tree, and a Random Forest Classifier. The hyperparameters used for each of these classifiers were left at the Weka defaults as the intention of these experiments was not to optimize nor tune a new social bot detection algorithm but instead to provide a reasonable estimate of an upper bound of performance that could be achieved with a subset of predictor variables.

The Naïve Bayes classifier applies Baye's theorem with strong independence assumptions between features. The Naïve Bayes classifier applied in this experiment estimated its parameters using maximum likelihood. The logistic regression used in this experiment performs similar to traditional logistic regressions whereby the model predicts $P(Y=1)$ as a function of X. The model tested employed instance weights (le Cessie, van Houwelingen, 1992).

The J48 (C4.5) algorithm builds decision trees by calculating and operationalizing information entropy. This algorithm derives a p-dimensional vector of attribute values or features and the class in which the sample falls to determine the optimal split (or branch) that will increase the accuracy of the model's predictions. The normalized information gain forms the basis for the splitting criterion. This gain in information is calculated from the difference in entropy. The C4.5 algorithm recurses through the predictive attributes with the highest normalized information gain to select the order in which attributes will partition. A Random Forest classifier acts as a meta estimator by fitting several decision tree classifiers on various sub-samples from a training data set.

Attributes. The attributes used to generate each of these machine learning algorithms included:

Total Number of Tweets Generated, within Twitter known as the "statuses count", refers to the number of Tweets (including reTweets) issued by the user.

Account Age is a value derived from the "created at" attribute which documents the UTC datetime that the user account was created on Twitter. For this analysis, January, 2022 was used as the end-date in calculating total Account Age.

Number of Followers tracks the total number of followers this account currently has.

Number of Following tracks the number of users this account is following.

Model Performance. Performance assessment of each trained model compares the overall accuracy of the models' predictions and their Kappa Statistics. Cohen's Kappa Statistic was included because the training and test data were imbalanced (2760 'bot's and 1689 'humans'). The Kappa Statistic classifier demonstrates performance over a classifier that simply guesses at random. The Kappa Statistic is always less than or equal to 1. We will interpret these values using Landis and Koch's approach (1977). Values < 0 will be interpreted as indicating no agreement, values ranging from 0–0.20 as slight, 0.21–0.40 as fair, 0.41–0.60 as moderate, 0.61–0.80 as substantial, and 0.81–1 as nearly perfect agreement.

Table 1H

Machine Learning Experiment 1 Results.

Algorithm	**%Correct**	**Kappa Statistic**
Naïve Bayes	75.16%	0.46
Logistic Regression	79.38%	0.56
J48 (C4.5) Decision Tree	84.03%	0.66
Random Forest	85.82%	0.70

Note. All settings are default unless otherwise indicated. Results of first experiment using attributes (1) account age, (2) total number of Tweets, (3) number of following, and (4) number of followers and ten-fold cross-validation in Weka.

Discussion. The accuracy levels of all four models outperformed chance. Furthermore, the Kappa Statistics for all models, particularly the decision tree and random forest classifiers, demonstrated substantial agreement. This finding indicates that sufficient information is available from the account age, the total number of Tweets, number of following, and number of followers, to improve social bot detection performance beyond chance.

Experiment 2

The data used to train the next set of models was the same as that used in the previous experiment. The second experiment compares the performance of the same classifiers with the same default hyperparameters. However, these classifiers are trained without access to the attributes used in the previous models. Instead, they will be trained using all the remaining Tier 1 attributes that do not indicate account age, number of Tweets, number following, or the number of followers or attributes such as tweets per day which be derived from these features.

Attributes. The attributes used to generate each of these machine learning algorithms included:

Screen Name is the handle, or alias with which users identify themselves.

Name Length is the count of characters contained within a user's name.

String Entropy is a measure of the random nature of the strings used in a user name.

Has Description. A binary 'yes' 'no' coding for whether a user has provided a user-defined UTF-8 string describing their account.

Source refers to the mode with which Twitter is interaction with (i.e., on an Android Device, Apple Device, iPhone, etc.

Has Location. A binary 'yes' 'no' coding for whether a user has provided a user-defined location for their account's profile.

Bot Reference is a binary 'yes' 'no' coding for whether the term 'bot' is found within the profile.

Favorites Count is the total number of likes an account has generated.

Status is ReTweet is a binary 'yes' 'no' coding for whether the last Tweet was a reTweet.

Last Status Hashtag is a count of the total number of hashtags in the last user Tweet.

Last Status Mentions is a count of the total number of other accounts tagged in the last user Tweet.

Status Possibly Sensitive is a binary 'yes' 'no' coding for whether the last tweet contains sensitive content.

Has Default Profile is a binary 'yes' 'no' coding for whether the profile has been update by the user or remained with default settings.

Emojis in Name is a count of the total emojis used within a user's name.

Emojis in Description is a count of the total emojis used within a user's description.

Verified. When true, it indicates that the user has a verified account. According to Twitter, "The blue Verified badge on Twitter lets people know that an account of public interest is authentic. To receive the blue badge, your account must be authentic, notable, and active."

Status ID refers is an integer representation of the unique identifier for a given User.

Table 2H

Machine Learning Experiment 2 Results.

Algorithm	%Correct	Kappa Statistic
Naïve Bayes	63.27%	0.30
Logistic Regression	67.73%	0.31
J48 (C4.5) Decision Tree	68.53%	0.34
Random Forest	68.40%	0.23

Note. All settings are default unless otherwise indicated. Results of second experiment using remaining attributes of Bot Hunter Tier 1 algorithm and ten-fold cross-validation in Weka.

Discussion. This experiment demonstrates a fall-off in performance relative to the performance of the models trained with a limited set of Twitter managed features in experiment 1. The accuracy of the models' predictions and their Kappa Statistics were lower than those observed in Experiment 1. This suggests that a training protocol centered on a limited set of platform-managed features can outperform models with access to more extensive features that primarily include user-defined attributes.

Experiment 3

The final experiment used the same data used to train the models in the first two experiments. It also compares the performance of the same classifiers used in the previous experiments, again with the same default hyperparameters. However, we test the performance of this model against the 60 stimuli used in study 2 to determine if they would be suitable for use in study 3. We will determine the suitability of these stimuli based on the consistency of the performance of these models relative to the test data.

Should the accuracy scores of these four models be close to or better than that observed performance in the previous experiments, we will conclude that the information within the 60 new stimuli would be appropriate to test participants' bot detection abilities with and without training.

Table 3H

Machine Learning Experiment 3 Results

Algorithm	Accuracy	Precision	Recall	ROC Area	PRC Area
Naïve Bayes	75%	0.764	0.750	0.874	0.864
Logistic Regression	78.33%	0.786	0.783	0.882	0.878
J48 (C4.5) Decision Tree	90%	0.902	0.900	0.912	0.888
Random Forest	90%	0.900	0.900	0.938	0.926

Note. All settings are default unless otherwise indicated. Results of third experiment using attributes (1) account age, (2) total number of Tweets, (3) number of following, and (4) number of followers and ten-fold cross-validation in Weka.

Discussion. As seen in the high accuracy levels of each of the classifiers, sufficient information relaying bot signals appears to be present in the 60 stimuli selected for use in study 3. The percent of correctly labeled personas within the 60 stimuli by the trained classifiers matched or surpassed the performance seen in Experiment 1 and reported in Table 1. The Decision Tree classifiers consistently outperformed both Naïve Bayes and Logistic Regression approaches in Experiments 1 and 3. We will incorporate this finding with our approach to training a social bot detection.

Appendix I

Confidence in Bot Aid Indicator Scores

In both studies 2 and 3, we predicted participants in the aid conditions would show signs of distrust in the bot indicator score as it approached maximum ambiguity, the 50% threshold. The following figures provide our findings from the aid group participants. The SDT responses were coded as binary 1 for 'bot' and 0 for 'human.' We calculated the following:

1. Proportion of 'bot' responses. The more participants in the aid group that responded 'bot' the higher this proportion.
2. Variance of the binary response. The more participants were divided in their views of whether a persona was a bot or a human the greater the variance (i.e., equal distribution of 1s and 0s). Increased variance would reflect a hesitancy in some participants to default to the bot indicator's recommendation
3. Mean of the confidence responses. Participants, after making their bot detection judgment, rated their confidence in that response from 50% to 100% confidence. Decreased confidence around the 50% threshold would mirror the rising uncertaint of the bot indicator score at this tipping point.

In the following figures, because the exact bot indicator scores for stimuli varied, they are arranged ordinally, from lowest bot probabilities the to highest.

Study 2

Figure 1I

Study 2 – Proportion 'Bot' Response

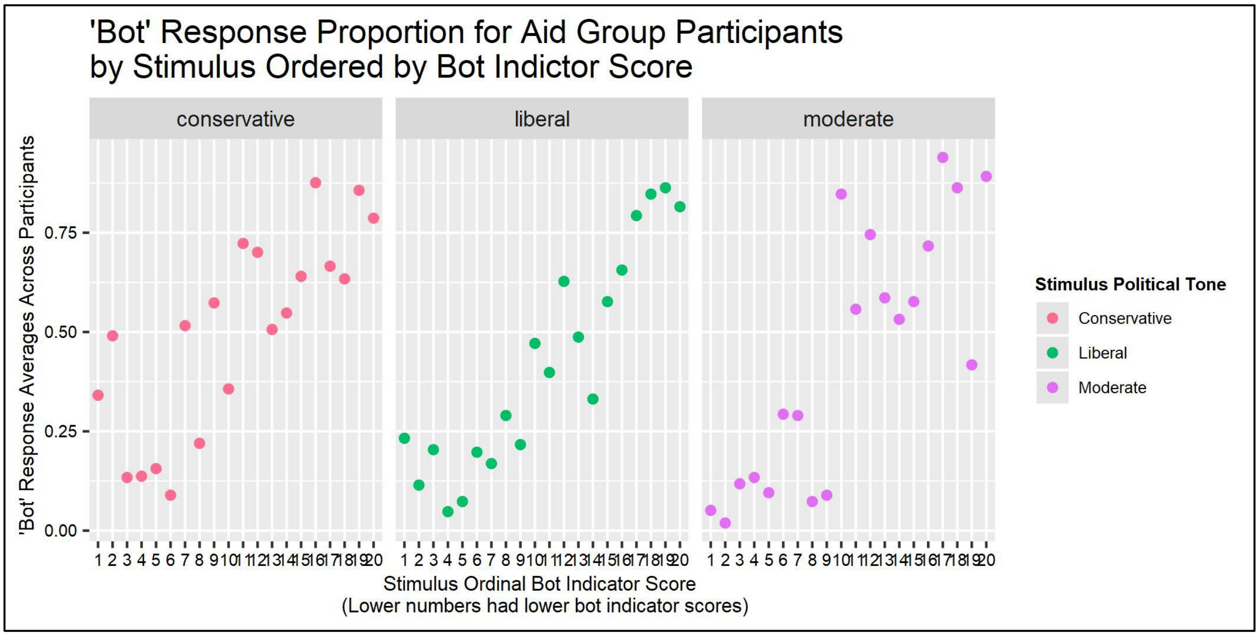


Figure 2I

Study 2 - Variance

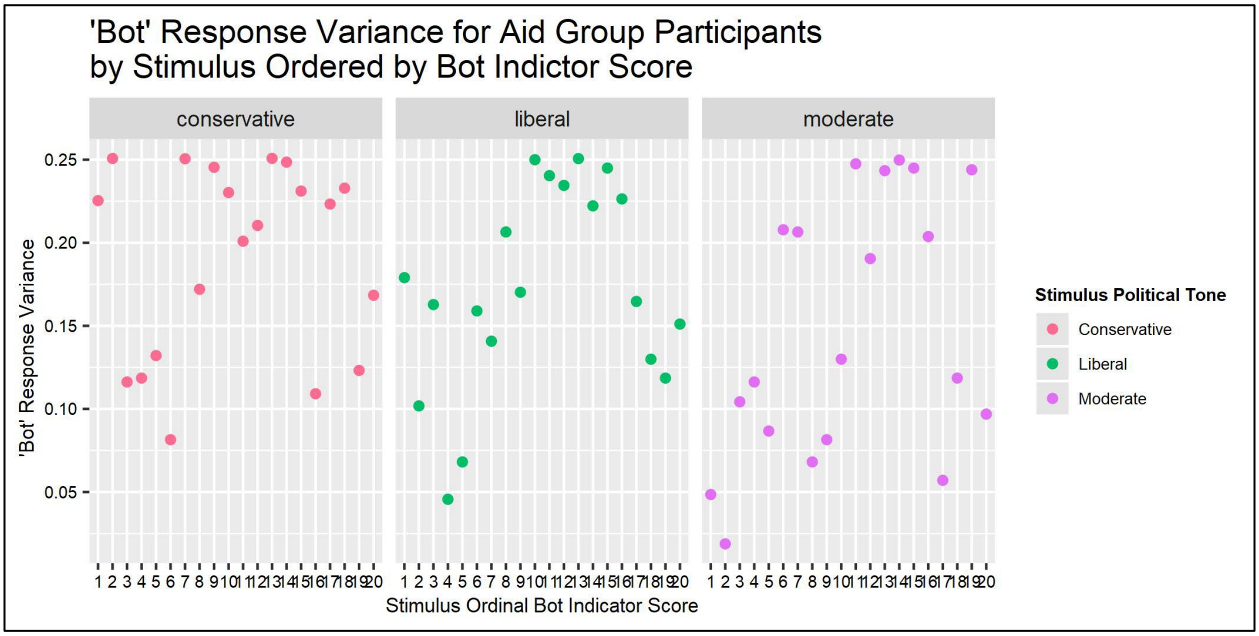
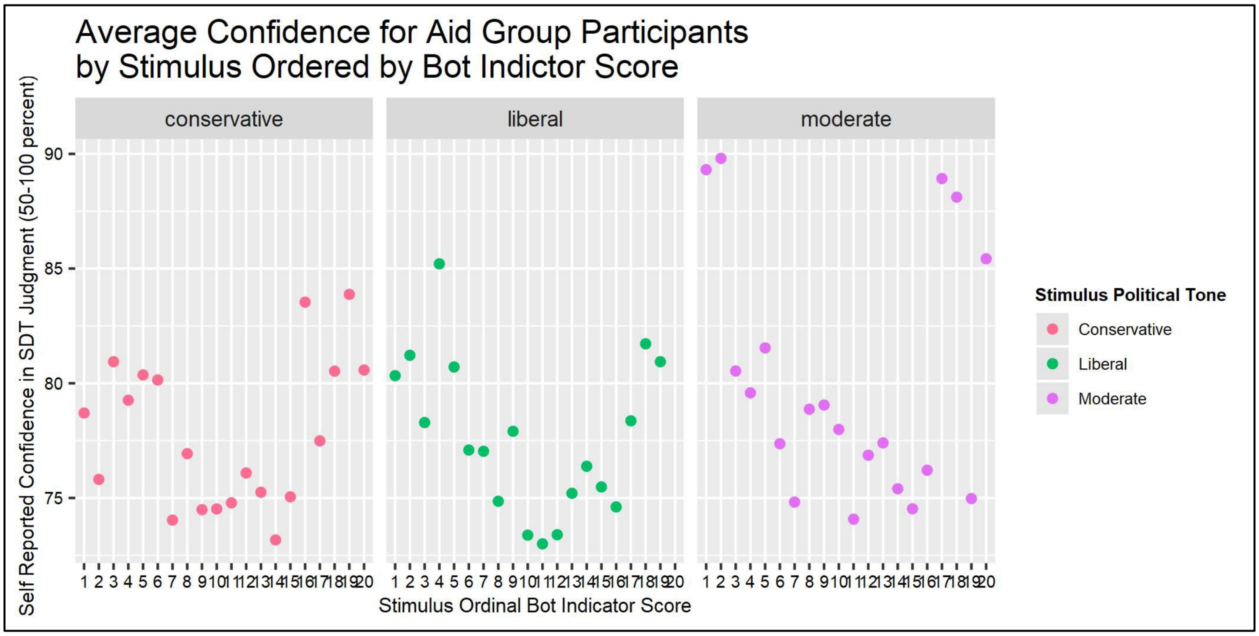


Figure 3I

Study 2 – Average Confidence



Study 3

Figure 4I

Study 3 – Proportion 'Bot' Response

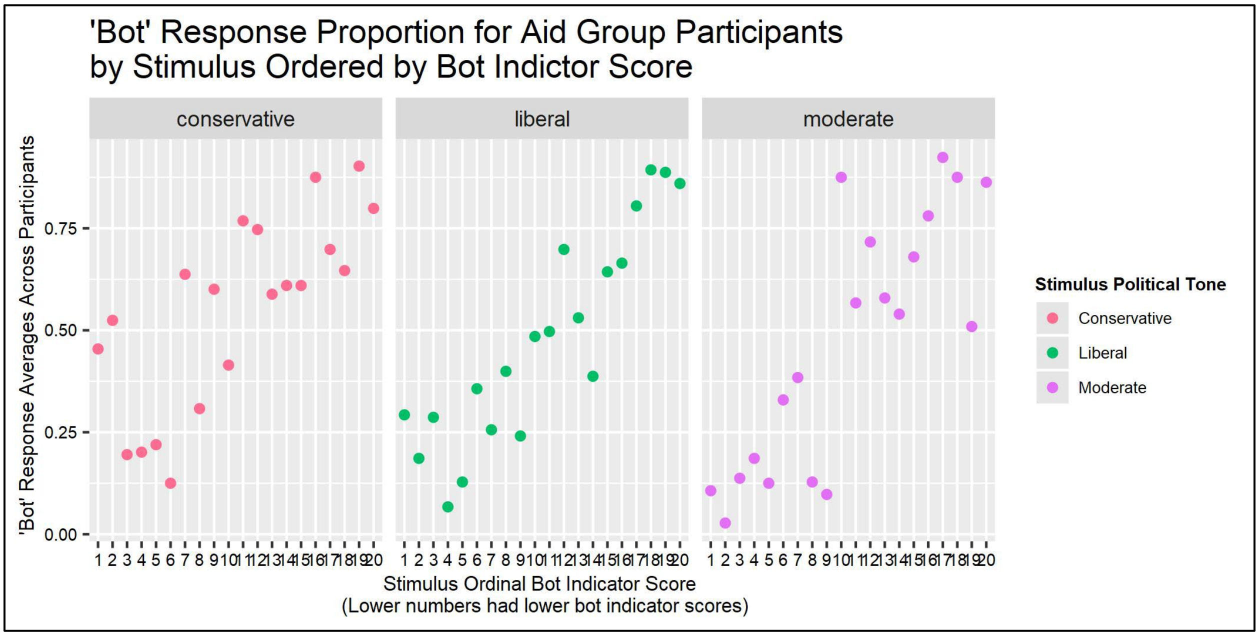


Figure 5I

Study 3 – Variance

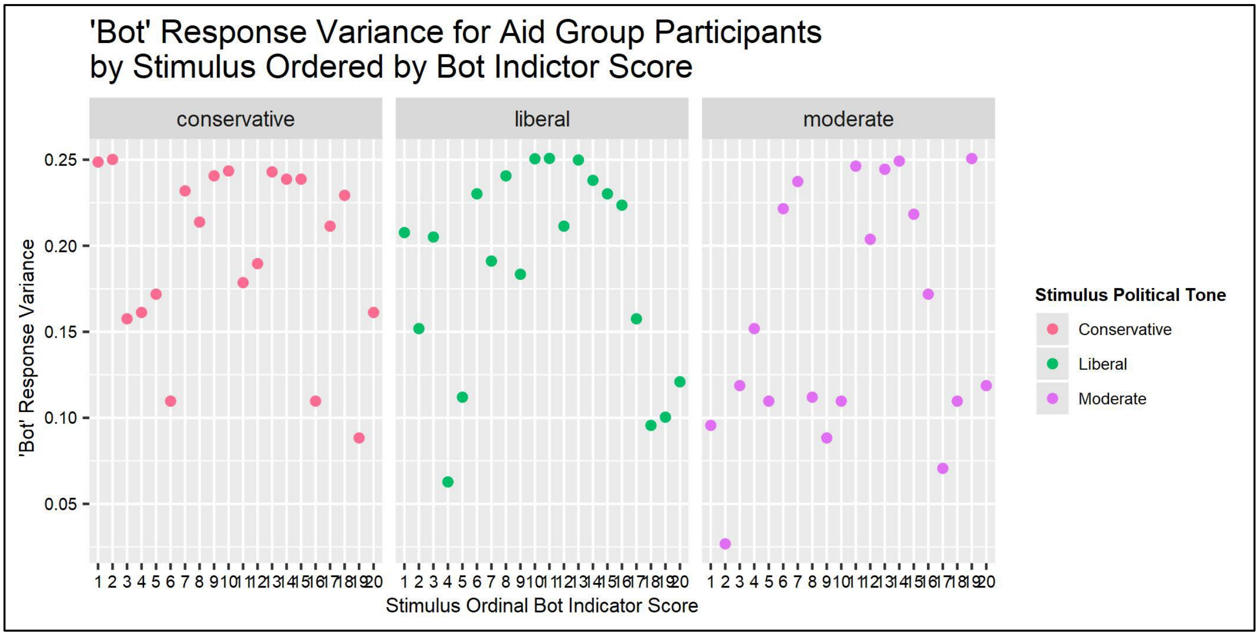
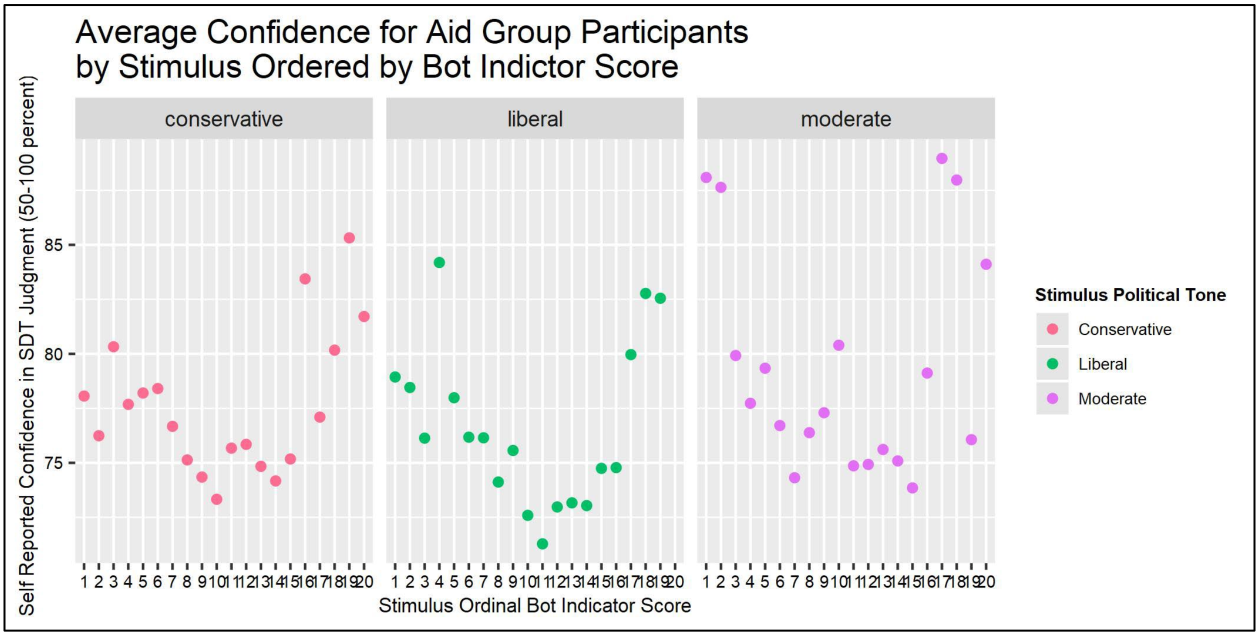


Figure 6I

Study 3 – Average Confidence



BIBLIOGRAPHY

Acquisti, A., Adjerid, I., Balebako, R., Brandimarte, L., Cranor, L. F., Komanduri, S., ... & Wilson, S. (2017). Nudges for privacy and security: Understanding and assisting users' choices online. ACM Computing Surveys (CSUR), 50(3), 1-41.

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE access, 6, 52138-52160.

Aiello, L. M., Deplano, M., Schifanella, R., & Ruffo, G. (2012). People are strange when you're a stranger: Impact and influence of bots on social networks. In Proceedings of the International AAAI Conference on Web and Social Media (Vol. 6, No. 1).

Allen, R., Mcgeorge, P., Pearson, D., & Milne, A. B. (2004). Attention and expertise in multiple target tracking. Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition, 18(3), 337-347.

Altay, S., Majima, Y., & Mercier, H. (2020). It's my idea! Reputation management and idea appropriation. Evolution and Human Behavior, 41(3), 235-243.

Altay, S., Hacquin, A. S., & Mercier, H. (2020). Why do so few people share fake news? It hurts their reputation. new media & society, 1461444820969893.

Andriosopoulos, D., Doumpos, M., Pardalos, P. M., & Zopounidis, C. (2019). Computational approaches and data analytics in financial services: A literature review. Journal of the Operational Research Society, 70(10), 1581-1599.

Anjomshoae, S., Najjar, A., Calvaresi, D., & Främling, K. (2019). Explainable agents and robots: Results from a systematic literature review. In 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019 (pp. 1078-1088). International Foundation for Autonomous Agents and Multiagent Systems.

Appel, G., Grewal, L., Hadi, R., & Stephen, A. T. (2020). The future of social media in marketing. Journal of the Academy of Marketing Science, 48(1), 79-95.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82-115.

Bakshy, E., Messing, S., & Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on Facebook. Science, 348(6239), 1130-1132.

Bar-Hillel, M., Noah, T., & Frederick, S. (2019.). Solving stumpers, CRT and CRAT: Are the abilities related? Judgment and Decision Making, 4.

Bahner, J. E., Hüper, A. D., & Manzey, D. (2008). Misuse of automated decision aids: Complacency, automation bias and the impact of training experience. International Journal of Human-Computer Studies, 66(9), 688-699.

Banker, S., & Khetani, S. (2019). Algorithm overdependence: How the use of algorithmic recommendation systems can increase risks to consumer well-being. Journal of Public Policy & Marketing, 38(4), 500-515.

Barberá, P., Jost, J. T., Nagler, J., Tucker, J. A., & Bonneau, R. (2015). Tweeting from left to right: Is online political communication more than an echo chamber?. Psychological science, 26(10), 1531-1542.

Baron, J. (2017). Comment on Kahan and Corbin: Can polarization increase with actively open-minded thinking?. Research & Politics, 4(1), 2053168016688122.

Baron, J. (2019). Actively open-minded thinking in politics. Cognition, 188, 8-18.

Baron, J., & Jost, J. T. (2019). False equivalence: Are liberals and conservatives in the United States equally biased?. Perspectives on Psychological Science, 14(2), 292-303.

Bastos, M. T., & Mercea, D. (2019). The Brexit botnet and user-generated hyperpartisan news. Social science computer review, 37(1), 38-54.

Beatson, O., Gibson, R., Cunill, M. C., & Elliot, M. (2021). Automation on twitter: Measuring the effectiveness of approaches to bot detection. Social Science Computer Review, 08944393211034991.

Beskow, D. (2020). Finding and Characterizing Information Warfare Campaigns. Doctoral Dissertation, Institute in Software Research, School of Computer Science, Carnegie Mellon University, http://reports-archive.adm.cs.cmu.edu/anon/isr2020/CMU-ISR-20-107.pdf.

Beskow, D. & Carley, K. (2018a). Bot-hunter: A tiered approach to detecting & characterizing automated activity on twitter. Conference: SBP-BRiMS: International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation, Washington DC.

Beskow, D. & Carley, K. (2018b). Introducing bothunter: A tiered approach to detection and characterizing automated activity on twitter. In Halil Bisgin, Ayaz Hyder, Chris Dancy, and Robert Thomson, editors, International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation. Springer, 2018.

Beskow, D. M., & Carley, K. M. (2018c). Bot conversations are different:

leveraging network metrics for bot detection in twitter. In 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM) (pp. 825-832). IEEE.

Benkler, Y., Faris, R., & Roberts, H. (2018). Network propaganda: Manipulation, disinformation, and radicalization in American politics. Oxford University Press.Bessi, A., & Ferrara, E. (2016). Social bots distort the 2016 US Presidential election online discussion. First monday, 21(11-7).

Bialek, M., & Pennycook, G. (2018). The cognitive reflection test is robust to multiple exposures. Behavior Research Methods, 50(5), 1953–1959. https://doi.org/10.3758/s13428-017-0963-x.

Bisseret, A. (1981). Application of signal detection theory to decision making in supervisory control: The effect of the operator's experience. Ergonomics, 24(2), 81-94.

Bond, G. D. (2008). Deception detection expertise. Law and Human Behavior, 32(4), 339-351.

Bond Jr, C. F., & DePaulo, B. M. (2006). Accuracy of deception judgments. Personality and social psychology Review, 10(3), 214-234.

Bond Jr, C. F., & DePaulo, B. M. (2008). Individual differences in judging deception: accuracy and bias. Psychological Bulletin, 134(4), 477.

Botometer. (2021). CNetS. Retrieved March 22, 2021, from https://cnets.indiana.edu/blog/tag/botometer/

Bot Sight (2022). https://www.nortonlifelock.com/blogs/norton-labs/botsight-tool-detect-twitter-bots.

Bovet, A., & Makse, H. A. (2019). Influence of fake news in Twitter during the 2016 US presidential election. Nature communications, 10(1), 1-14.

Brady, W. J., Wills, J. A., Burkart, D., Jost, J. T., & Van Bavel, J. J. (2019). An asymmetry in the diffusion of moralized content on social media among political leaders. Journal of Experimental Psychology: General, 148(10), 1802.

Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). Statistical science, 16(3), 199-231.

Bronstein, M. V., Pennycook, G., Bear, A., Rand, D. G., & Cannon, T. D. (2019). Belief in Fake news is associated with delusionality, dogmatism, religious fundamentalism, and reduced analytic thinking. Journal of Applied Research in Memory and Cognition, 8(1), 108-117.

Buck, C., Doctor, E., Hennrich, J., Jöhnk, J., & Eymann, T. (2022). General Practitioners' Attitudes Toward Artificial Intelligence–Enabled Systems: Interview Study. Journal of Medical Internet Research, 24(1), e28916.

Cai, C. W. (2020). Nudging the financial market? A review of the nudge theory. Accounting & Finance, 60(4), 3341-3365.

Caldarelli, G., De Nicola, R., Del Vigna, F., Petrocchi, M., & Saracco, F. (2020). The role of bot squads in the political propaganda on Twitter. Communications Physics, 3(1), 1-15.

Caldarelli, G., De Nicola, R., Petrocchi, M., Pratelli, M., & Saracco, F. (2021). Flow of online misinformation during the peak of the COVID-19 pandemic in Italy. EPJ data science, 10(1), 34.

Campitelli, G., & Labollita, M. (2010). Correlations of cognitive reflection with judgments and choices. Judgment and Decision making, 5(3), 182-191.

Canfield, C. I., Fischhoff, B., & Davis, A. (2016). Quantifying phishing susceptibility for detection and behavior decisions. Human factors, 58(8), 1158-1172.

Cañal-Bruland, R., & Schmidt, M. (2009). Response bias in judging deceptive movements. Acta psychologica, 130(3), 235-240.

Choi, D., Chun, S., Oh, H., & Han, J. (2020). Rumor propagation is amplified by echo chambers in social media. Scientific reports, 10(1), 1-10.

Chu, Z., Gianvecchio, S., Wang, H., & Jajodia, S. (2010). Who is tweeting on Twitter: human, bot, or cyborg?. In Proceedings of the 26th annual computer security applications conference (pp. 21-30).

Cinelli, M., Morales, G. D. F., Galeazzi, A., Quattrociocchi, W., & Starnini, M. (2021). The echo chamber effect on social media. Proceedings of the National Academy of Sciences, 118(9).

Clemons, E. K. (2009). The complex problem of monetizing virtual electronic social networks. Decision support systems, 48(1), 46-56.

Colleoni, E., Rozza, A., & Arvidsson, A. (2014). Echo chamber or public sphere? Predicting political orientation and measuring political homophily in Twitter using big data. Journal of communication, 64(2), 317-332.

Collier, C. A. (2018, July). Nudge Theory in Information Systems Research A Comprehensive Systematic Review of the Literature. In Academy of Management Proceedings (Vol. 2018, No. 1, p. 18642). Briarcliff Manor, NY 10510: Academy of Management.

Constantine, L. (2022). "How data poisoning attacks corrupt machine learning models," CSO, retrieved from: https://www.csoonline.com/article/3613932/how-data-poisoning-attacks-corrupt-machine-learning-models.html, 14 March 2022.

Cook, D., Waugh, B., Abdipanah, M., Hashemi, O., & Rahman, S. (2014). Twitter Deception and Influence: Issues of Identity, Slacktivism, and Puppetry. Journal of Information Warfare, 13(1), 58-71. Retrieved November 3, 2020, from https://www.jstor.org/stable/26487011.

Coppock, A., & McClellan, O. A. (2019). Validating the demographic, political, psychological, and experimental results obtained from a new source of online survey respondents. Research & Politics, 6(1), 2053168018822174.

Cresci, S. (2019). Detecting malicious social bots: story of a never-ending clash. In Multidisciplinary International Symposium on Disinformation in Open Online Media (pp. 77-88). Springer, Cham.

Cresci, S. (2020). A decade of social bot detection. Communications of the ACM, 63(10), 72-83.

Cresci, S., Petrocchi, M., Spognardi, A., & Tognazzi, S. (2021). The coming age of adversarial social bot detection. First Monday.

Crump, M. J., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. PloS one, 8(3), e57410.

Dahmani, L., & Bohbot, V. D. (2020). Habitual use of GPS negatively impacts spatial memory during self-guided navigation. Scientific reports, 10(1), 1-14.

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

Davis, C. A., Varol, O., Ferrara, E., Flammini, A., & Menczer, F. (2016). BotOrNot: A system to evaluate social bots. Proceedings of the 25th International Conference Companion on World Wide Web - WWW '16 Companion, 273–274. https://doi.org/10.1145/2872518.2889302

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. American psychologist, 34(7), 571.

DeCarlo, L. T. (1998). Signal detection theory and generalized linear models. Psychological methods, 3(2), 186.

Deppe, K. D., Gonzalez, F. J., Neiman, J., Pahlke, J., Smith, K., & Hibbing, J. R. (2015). Reflective liberals and intuitive conservatives: A look at the Cognitive Reflection Test and ideology.

Dewitt, B., Fischhoff, B., Davis, A., & Broomell, S. B. (2015). Environmental risk perception from visual cues: The psychophysics of tornado risk perception. Environmental Research Letters, 10(12), 124009.

Diab, D. L., Pui, S. Y., Yankelevich, M., & Highhouse, S. (2011). Lay perceptions of selection decision aids in US and non-US samples. International Journal of Selection and Assessment, 19(2), 209-216.

Diehl, R. L. (1981). "Feature detectors for speech: A critical reappraisal." Psychological Bulletin 89.1:1.

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.

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.

Ditto, P. H., Liu, B. S., Clark, C. J., Wojcik, S. P., Chen, E. E., Grady, R. H., ... & Zinger, J. F. (2019). At least bias is bipartisan: A meta-analytic comparison of partisan bias in liberals and conservatives. Perspectives on Psychological Science, 14(2), 273-291.

Downs, J. S., Holbrook, M. B., Sheng, S., & Cranor, L. F. (2010). Are your participants gaming the system? Screening Mechanical Turk workers. In Proceedings of the SIGCHI conference on human factors in computing systems (pp. 2399-2402).

Drummond, C., & Fischhoff, B. (2017). Individuals with greater science literacy and education have more polarized beliefs on controversial science topics. PNAS, 114, 9587-9592

Drummond, C., & Fischhoff, B. (2019). Does “putting on your thinking cap” reduce myside bias in evaluation of scientific evidence? Thinking & Reasoning, 25(4), 477–505. https://doi.org/10.1080/13546783.2018.1548379

Dunning, D., Meyerowitz, J. A., & Holzberg, A. D. (1989). Ambiguity and self-evaluation: The role of idiosyncratic trait definitions in self-serving assessments of ability. Journal of personality and social psychology, 57(6), 1082.

Dwivedi, Y. K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., ... & Williams, M. D. (2021). Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. International Journal of Information Management, 57, 101994.

Endsley, M. R. (2018). Expertise and situation awareness. In K. A. Ericsson, R. R. Hoffman, A. Kozbelt, A. M. Williams (Eds.), The Cambridge Handbook of Expertise and Expert Performance (2nd ed., pp. 714–742). Cambridge University Press. https://doi.org/10.1017/9781316480748.037.

Eibe Frank, Mark A. Hall, and Ian H. Witten (2016). The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016.

Ericsson, K. A. (2017). Expertise and individual differences: The search for the structure and acquisition of experts' superior performance. Wiley Interdisciplinary Reviews: Cognitive Science, 8(1-2), e1382.

Ericsson, K. A. (2018). The differential influence of experience, practice, and deliberate practice on the development of superior individual performance of experts. In K. A. Ericsson, R. R. Hoffman, A. Kozbelt, & A. M. Williams (Eds.), Cambridge handbooks in psychology. The Cambridge handbook of expertise and expert performance (p. 745–769). Cambridge University Press. https://doi.org/10.1017/9781316480748.038

Ericsson, K.A., Krampe, R.T., Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. Psychol Rev, 100:363–406.

Ericsson, A., & Pool, R. (2016). Peak: Secrets from the new science of expertise. Random House.

Eyal, P., David, R., Andrew, G., Zak, E., & Ekaterina, D. (2021). Data quality of platforms and panels for online behavioral research. Behavior Research Methods, 1-20.

Ferrara, E., Varol, O., Davis, C., Menczer, F., Flammini, A. (2016). The rise of social bots. Communications of the ACM 59(7), 96–104.

Farewell, V. T., Long, D. L., Tom, B. D. M., Yiu, S., & Su, L. (2017). Two-part and related regression models for longitudinal data. Annual review of statistics and its application, 4, 283-315.

Felton, M., Crowell, A., & Liu, T. (2015). Arguing to agree: Mitigating my-side bias through consensus-seeking dialogue. Written Communication, 32(3), 317-331.

Ferrara, E., Cresci, S., & Luceri, L. (2020). Misinformation, manipulation, and abuse on social media in the era of COVID-19. Journal of Computational Social Science, 3(2), 271-277.

Ferrara, E., Varol, O., Davis, C., Menczer, F., Flammini, A. (2016). The rise of social bots. Communications of the ACM, 59(7), 96–104. https://doi.org/10.1145/2818717.

Flaxman, S., Goel, S., & Rao, J. M. (2016). Filter bubbles, echo chambers, and online news consumption. Public opinion quarterly, 80(S1), 298-320.

Fogg, B. J. (2002). Persuasive technology: using computers to change what we think and do. Ubiquity, 2002(December), 2.

Frederick, S. (2005). Cognitive Reflection and Decision Making. Journal of Economic Perspectives, 19(4), 25–42. https://doi.org/10.1257/089533005775196732

Freelon, D., Marwick, A., & Kreiss, D. (2020). False equivalencies: Online activism from left to right. Science, 369(6508), 1197-1201.

Frenda, S. J., Knowles, E. D., Saletan, W., & Loftus, E. F. (2013). False memories of fabricated political events. Journal of Experimental Social Psychology, 49(2), 280-286.

Gaube, S., Suresh, H., Raue, M., Merritt, A., Berkowitz, S. J., Lermer, E., ... & Ghassemi, M. (2021). Do as AI say: susceptibility in deployment of clinical decision-aids. NPJ digital medicine, 4(1), 1-8.

Gelman, A., Su, Y. S., Yajima, M., Hill, J., Pittau, M. G., Kerman, J., & Dorie, V. (2016). Package ‘Arm’: Data Analysis Using Regression and Multilevel/Hierarchical Models. R Package Version 1.9-3.

Ghazizadeh, M., Lee, J. D., & Boyle, L. N. (2012). Extending the Technology Acceptance Model to assess automation. Cognition, Technology & Work, 14(1), 39-49.

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. Academy of Management Annals, 14(2), 627-660.

Gramlich, J. (2020). Democrats, Republicans each expect made-up news to target their own party more than the other in 2020. Retrieved April 5, 2021, from https://www.pewresearch.org/fact-tank/2020/02/11/democrats-republicans-each-expect-made-up-news-to-target-their-own-party-more-than-the-other-in-2020/

Green, D. M., & Swets, J. A. (1966). Signal detection theory and psychophysics (Vol. 1, pp. 1969-12). New York: Wiley.

Guo, L., A. Rohde, J., & Wu, H. D. (2020). Who is responsible for Twitter's echo chamber problem? Evidence from 2016 US election networks. Information, Communication & Society, 23(2), 234-251.

Haidt, J. (2012). The righteous mind: Why good people are divided by politics and religion. Vintage.

Hajli, N., Saeed, U., Tajvidi, M., & Shirazi, F. (2021). Social Bots and the Spread of

Disinformation in Social Media: The Challenges of Artificial Intelligence. British Journal of Management.

Hanushek, E. A., & Rivkin, S. G. (2010). Generalizations about using value-added measures of teacher quality. American Economic Review, 100(2), 267-71.

Hartwig, M., Granhag, P. A., Strömwall, L. A., & Kronkvist, O. (2006). Strategic use of evidence during police interviews: When training to detect deception works. Law and human behavior, 30(5), 603.

Hauch, V., Sporer, S. L., Michael, S. W., & Meissner, C. A. (2016). Does training improve the detection of deception? A meta-analysis. Communication Research, 43(3), 283-343.

Hauser, D. J., & Schwarz, N. (2016). Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. Behavior research methods, 48(1), 400-407.

Havey, N. F. (2020). Partisan public health: How does political ideology influence support for COVID-19 related misinformation?. Journal of Computational Social Science, 3(2), 319-342.

Hayes, J. L., Brinson, N. H., Bott, G. J., & Moeller, C. M. (2021). The Influence of Consumer–Brand Relationship on the Personalized Advertising Privacy Calculus in Social Media. Journal of Interactive Marketing, 55, 16-30.

Holzinger, A., Kieseberg, P., Weippl, E., & Tjoa, A. M. (2018). Current advances, trends and challenges of machine learning and knowledge extraction: from machine learning to explainable AI. In International Cross-Domain Conference for Machine

Learning and Knowledge Extraction (pp. 1-8). Springer, Cham.

Hjorth, F., & Adler-Nissen, R. (2019). Ideological asymmetry in the reach of pro-Russian digital disinformation to United States audiences. Journal of Communication, 69(2), 168-192.

Hjouji, Z.e., Hunter, D.S., Mesnards, N.G.d., Zaman, T. (2018). The impact of bots on opinions in social networks. arXiv preprint arXiv:1810.12398.

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. Human factors, 57(3), 407-434.

Hou, S.-I. (2017). Measuring social media active level (SMACTIVE) and engagement level (SMENGAGE) among professionals in higher education. International Journal of Cyber Society and Education, 10(1), 1–16. https://doi.org/10.7903/ijcse.1520

Huang, B. (2020). "Learning User Latent Attributes on Social Media," Ph.D. Thesis, School of Computer Science, Institute of Software Research, Carnegie Mellon University, Pittsburgh, PA, USA.

Huang, B., & Carley, K. M. (2020). Disinformation and misinformation on twitter during the novel coronavirus outbreak. arXiv preprint arXiv:2006.04278.

Iyengar, S., & Hahn, K. S. (2009). Red media, blue media: Evidence of ideological selectivity in media use. Journal of communication, 59(1), 19-39.

Iyengar, S., Lelkes, Y., Levendusky, M., Malhotra, N., & Westwood, S. J. (2019). The origins and consequences of affective polarization in the United States. Annual Review of Political Science, 22, 129-146.

Jagielski, M., Oprea, A., Biggio, B., Liu, C., Nita-Rotaru, C., & Li, B. (2018). Manipulating machine learning: Poisoning attacks and countermeasures for regression learning. In 2018 IEEE Symposium on Security and Privacy (SP) (pp. 19-35). IEEE.

Jansen, B. J., Zhang, M., Sobel, K., & Chowdury, A. (2009). Twitter power: Tweets as electronic word of mouth. Journal of the American society for information science and technology, 60(11), 2169-2188.

Jost, J. T. (2017). Ideological asymmetries and the essence of political psychology. Political psychology, 38(2), 167-208.

Jost, J. T., Glaser, J., Kruglanski, A. W., & Sulloway, F. J. (2003). Political conservatism as motivated social cognition. Psychological bulletin, 129(3), 339.

Jost, J. T., Stern, C., Rule, N. O., & Sterling, J. (2017). The politics of fear: Is there an ideological asymmetry in existential motivation?. Social cognition, 35(4), 324-353.

Jost, J. T., van der Linden, S., Panagopoulos, C., & Hardin, C. D. (2018). Ideological asymmetries in conformity, desire for shared reality, and the spread of misinformation. Current opinion in psychology, 23, 77-83.

Jurkovič, M. (2016). Effect of short-term mindfulness induction on myside bias and miserly processing: A preliminary study. Studia Psychologica, 58(3), 231-237.

Kahn, K. B., & Davies, P. G. (2011). Differentially dangerous? Phenotypic racial stereotypicality increases implicit bias among ingroup and outgroup members. Group Processes & Intergroup Relations, 14(4), 569-580.

Kahan, D. M. (2012). Ideology, motivated reasoning, and cognitive reflection: An experimental study. Judgment and Decision making, 8, 407-24.

Kahan, D. M., Peters, E., Dawson, E. C., & Slovic, P. (2013). Motivated numeracy and enlightened self-government. Yale Law School Public Law & Legal Theory.

Public Working Paper, 116. Retrieved from http://papers. ssrn. com/sol3/Delivery. cfm/SSRN_ ID2319992_code45442. pdf.

Kahan, D. M., Landrum, A., Carpenter, K., Helft, L., & Hall Jamieson, K. (2017). Science curiosity and political information processing. Political Psychology, 38, 179-199.

Kahneman, D. & Frederick, S. (2002). Representativeness revisited: attribute substitution in intuitive judgment. In: Gilgovich T, Griffin D, Kahneman D, eds. Heuristics and Biases: the Psychology of Intuitive Judgment. New York: Cambridge University Press 2002.

Kahneman, D., & Tversky, A. (1972). Subjective Probability: A Judgment of Representativeness. Cognitive Psychology, 3:430–54.

Kahneman, D., & Klein, G. (2009). Conditions for intuitive expertise: a failure to disagree. American psychologist, 64(6), 515.

Kaivanto, K. (2014). The effect of decentralized behavioral decision making on system-level risk. Risk Analysis, 34(12), 2121-2142.

Kaivanto, K., Kroll, E. B., & Zabinski, M. (2014). Bias-trigger manipulation and task-form understanding in Monty Hall. Economics Bulletin, 34(1), 89-98.

Kats, D., (2022). Introducing BotSight: A new tool to detect bots on Twitter in real-time. NortonLifeLock. Retrieved February 21, 2022, from https://www.nortonlifelock.com/blogs/norton-labs/botsight-tool-detect-twitter-bots

Kenny, R., Fischhoff, B., Davis, A., Carley, K. M., & Canfield, C. (2022). Duped by bots: why some are better than others at detecting fake social media personas. Human factors, 00187208211072642.

Klayman, J. (1995). Varieties of confirmation bias. Psychology of learning and motivation, 32, 385-418.

Kluttz, D. N., & Mulligan, D. K. (2019). Automated decision support technologies and the legal profession. Berkeley Tech. LJ, 34, 853.

Kunda, Z. (1990). The case for motivated reasoning. Psychological bulletin, 108(3), 480.

Kurakin, A., Goodfellow, I., & Bengio, S. (2016). Adversarial machine learning at scale. arXiv preprint arXiv:1611.01236.

Landis, J.R.; Koch, G.G. (1977). "The measurement of observer agreement for categorical data". Biometrics 33 (1): 159–174

Landy, D. (2018). Perception in expertise. In K. A. Ericsson, R. R. Hoffman, A. Kozbelt, & A. M. Williams (Eds.), The Cambridge Handbook of Expertise and Expert Performance (2nd ed., pp. 151–164). Cambridge University Press. https://doi.org/10.1017/9781316480748.010

Langley, P. (1985). Learning to search: From weak methods to domain-specific heuristics. Cognitive Science, 9, 217-260.

Lee, C., Kwak, H., Park, H., & Moon, S. (2010). Finding influentials based on the temporal order of information adoption in twitter. In Proceedings of the 19th international conference on World wide web (pp. 1137-1138) for detailed perspectives on influence in social media environments.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance.

Human factors, 46(1), 50-80.

Lee, K., Eoff, B., & Caverlee, J. (2011). Seven months with the devils: A long-term study of content polluters on twitter. In Proceedings of the International AAAI Conference on Web and Social Media (Vol. 5, No. 1).

Lee, T., & Hosam, C. (2020). Fake news is real: the significance and sources of disbelief in mainstream media in Trump's America. In Sociological Forum (Vol. 35, pp. 996-1018).

Levay, K. E., Freese, J., & Druckman, J. N. (2016). The demographic and political composition of Mechanical Turk samples. Sage Open, 6(1), 2158244016636433.

Li, M., & Chapman, G. B. (2013). Nudge to health: Harnessing decision research to promote health behavior. Social and Personality Psychology Compass, 7(3), 187-198.

Liberman, A. M., Harris, K. S., Hoffman, H. S., & Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. Journal of experimental psychology, 54(5), 358.

Loepp, E., & Kelly, J. T. (2020). Distinction without a difference? An assessment of MTurk Worker types. Research & Politics, 7(1), 2053168019901185.

Longoni, C., Bonezzi, A., & Morewedge, C. K. (2019). Resistance to medical artificial intelligence. Journal of Consumer Research, 46(4), 629-650.

Loor, M., & De Tré, G. (2020). Contextualizing Naive Bayes Predictions. In International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (pp. 814-827). Springer, Cham.

Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. Journal of personality and social psychology, 37(11), 2098.

Lynn, S. K., & Barrett, L. F. (2014). "Utilizing" signal detection theory. Psychological science, 25(9), 1663-1673.

le Cessie, S., van Houwelingen, J.C. (1992). Ridge Estimators in Logistic Regression. Applied Statistics. 41(1):191-201.

Macmillan, N. A., & Creelman, C. D. (2004). Detection theory: A user's guide. Psychology press.

Mason, W., & Suri, S. (2012). Conducting behavioral research on Amazon’s Mechanical Turk. Behavior research methods, 44(1), 1-23.

Matthews, G., Warm, J. S., Reinerman, L. E., Langheim, L. K., & Saxby, D. J. (2010). Task engagement, attention, and executive control. In Handbook of individual differences in cognition (pp. 205-230). Springer, New York, NY.

McNicol, D. (2005). A primer of signal detection theory. Psychology Press.

Meehl, P. E. (1954). Clinical versus statistical prediction: A theoretical analysis and a review of the evidence.

Meyer, A., Zhou, E., Frederick, S., & Ackerman, R. (2018). The non-effects of repeated exposure to the Cognitive Reflection Test. Judgment and Decision Making, Vol. 13(3). https://doi.org/10.1037/xge0000049.

Mitchell, T., (1997). Machine Learning. McGraw-Hill, M. L. Edition.

Mohan, D., Angus, D. C., Ricketts, D., Farris, C., Fischhoff, B., Rosengart, M. R., ... & Barnato, A. E. (2014). Assessing the validity of using serious game technology to analyze physician decision making. PloS one, 9(8), e105445.

Mohan, D., Farris, C., Fischhoff, B., Rosengart, M. R., Angus, D. C., Yealy, D. M., ... & Barnato, A. E. (2017). Efficacy of educational video game versus traditional educational apps at improving physician decision making in trauma triage: randomized controlled trial. bmj, 359.

Mønsted, B., Sapieżyński, P., Ferrara, E., & Lehmann, S. (2017). Evidence of complex contagion of information in social media: An experiment using Twitter bots. PloS one, 12(9), e0184148.

Nickerson, R. S. (1998). Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. Review of General Psychology, 2(2), 175–220. https://doi.org/10.1037/1089-2680.2.2.175

Nikolov, D., Flammini, A., & Menczer, F. (2021). Right and left, partisanship predicts (asymmetric) vulnerability to misinformation. Harvard Kennedy School (HKS) Misinformation Review, 1(7).

Ohlsson, S. (1996). Learning from performance errors. Psychological review, 103(2), 241.

Önkal, D., Goodwin, P., Thomson, M., Gönül, S., & Pollock, A. (2009). The relative influence of advice from human experts and statistical methods on forecast adjustments. Journal of Behavioral Decision Making, 22(4), 390-409.

Orabi, M., Mouheb, D., Al Aghbari, Z., & Kamel, I. (2020). Detection of bots in social media: A systematic review. Information Processing & Management, 57(4), 102250.

Pacheco, D., Hui, P. M., Torres-Lugo, C., Truong, B. T., Flammini, A., & Menczer, F.

(2020). Uncovering coordinated networks on social media. arXiv preprint arXiv:2001.05658, 16.

Palan, S., & Schitter, C. (2018). Prolific. ac—A subject pool for online experiments. Journal of Behavioral and Experimental Finance, 17, 22-27.

Parasuraman, R., & Davies, D. R. (1977). A taxonomic analysis of vigilance performance. In vigilance (pp. 559-574). Springer, Boston, MA.

Pavlou, P. A. (2003). Consumer acceptance of electronic commerce: Integrating trust and risk with the technology acceptance model. International journal of electronic commerce, 7(3), 101-134.

Pe'er, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017). Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. Journal of Experimental Social Psychology, 70, 153-163.

Pe'er, E., Egelman, S., Harbach, M., Malkin, N., Mathur, A., & Frik, A. (2020). Nudge me right: Personalizing online security nudges to people's decision-making styles. Computers in Human Behavior, 109, 106347.

Pe'er, E., Rothschild, D. M., Evernden, Z., Gordon, A., Damer, E. (2021). Data Quality of Platforms and Panels for Online Behavioral Research. http://link.springer.com/article/10.3758/s13428-021-01694-3, Available at SSRN: https://ssrn.com/abstract=3765448 or http://dx.doi.org/10.2139/ssrn.3765448

Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A. A., Eckles, D., & Rand, D. G. (2021). Shifting attention to accuracy can reduce misinformation online. Nature, 592(7855), 590-595.

Pennycook, G., McPhetres, J., Zhang, Y., Lu, J. G., & Rand, D. G. (2020). Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy nudge intervention. Psychological science, 31(7), 770-780.

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

Pennycook, G., & Rand, D. (2021). Nudging social media sharing towards accuracy. forthcoming in The Annals of the American Academy of Political and Social Science, retrieved from: file:///C:/Users/ryank/Downloads/Accuracy%20prompt%20review_ANNALS_psyarxiv.pdf, 21 February 2022.

Pressley, M., Borkowski, J. G., & Schneider, W. (1987). Cognitive strategies: Good strategy users coordinate metacognition and knowledge.

Pronin, E., Gilovich, T., & Ross, L. (2004). Objectivity in the eye of the beholder: divergent perceptions of bias in self versus others. Psychological review, 111(3), 781.

Qiu, X., FM Oliveira, D., Sahami Shirazi, A., Flammini, A., & Menczer, F. (2017). Limited individual attention and online virality of low-quality information. Nature Human Behaviour, 1(7), 1-7.

Quigley, M. (2013). Nudging for health: on public policy and designing choice architecture. Medical law review, 21(4), 588-621.

Reeves, B., & Nass, C. (1996). The media equation: How people treat computers, television, and new media like real people. Cambridge, United Kingdom: Cambridge university press.

Rikers, R. M., Schmidt, H. G., & Boshuizen, H. P. (2000). Knowledge encapsulation and the intermediate effect. Contemporary educational psychology, 25(2), 150-166.

Riquelme, F., & González-Cantergiani, P. (2016). Measuring user influence on Twitter: A survey. Information processing & management, 52(5), 949-975.

Ross, R. M., Rand, D. G., & Pennycook, G. (2021). Beyond "fake news": Analytic thinking and the detection of false and hyperpartisan news headlines. Judgment and Decision Making, 16(2), 484-504.

Roozenbeek, J., Freeman, A. L., & van der Linden, S. (2021). How accurate are accuracy-nudge interventions? A preregistered direct replication of Pennycook et al.(2020). Psychological science, 32(7), 1169-1178.

Samuel, A.L. (1959). Some studies in machine learning using the game of checkers. IBM J. Res. Dev. 3, 3 (July), 210–229. DOI:https://doi.org/10.1147/rd.33.0210

Sarkar, S. (2016). "Accuracy and interpretability trade-offs in machine learning applied to safer gambling," in Proc. CEUR Workshop, pp. 79–87

Sarno, D. M., McPherson, R., & Neider, M. B. (2022). Is the key to phishing training persistence?: Developing a novel persistent intervention. Journal of Experimental Psychology: Applied.

Sasahara, K., Chen, W., Peng, H., Ciampaglia, G. L., Flammini, A., & Menczer, F. (2019). On the inevitability of online echo chambers. arXiv preprint arXiv:1905.03919, for a discussion as to why echo-chambers appear as emergent phenomenon.

Sembroski, C. E., Fraune, M. R., & Šabanović, S. (2017). He said, she said, it said: Effects of robot group membership and human authority on people's willingness to follow their instructions. In 2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN) (pp. 56-61). IEEE.

Shao, C., Ciampaglia, G. L., Varol, O., Yang, K.-C., Flammini, A., & Menczer, F. (2018). The spread of low-credibility content by social bots. Nature Communications, 9(1), 4787. https://doi.org/10.1038/s41467-018-06930-7.

Shahi, G. K., Dirkson, A., & Majchrzak, T. A. (2021). An exploratory study of covid-19 misinformation on twitter. Online social networks and media, 22, 100104.

Shorey, S., & Howard, P. N. (2016). Automation, big data, and politics: A research review. International Journal of Communication 10, 5032–5055.

Siau, K., & Wang, W. (2018). Building trust in artificial intelligence, machine learning, and robotics. Cutter Business Technology Journal, 31(2), 47–53. https://www.cutter.com/article/building-trustartificial-intelligence-machine-learning-and-robotics-498981

Simon, H.A., & Chase, W. G., (1973) Skill in chess. Am Sci, 61:394–403.

Sinclair, A. H., Stanley, M. L., & Seli, P. (2020). Closed-minded cognition: Right-wing authoritarianism is negatively related to belief updating following prediction error. Psychonomic bulletin & review, 27(6), 1348-1361.

Sorkin, R. D., & Woods, D. D. (1985). Systems with human monitors: A signal detection analysis. Human-computer interaction, 1(1), 49-75.

Spengler, P. M., White, M. J., Ægisdóttir, S., Maugherman, A. S., Anderson, L. A., Cook, R. S., ... & Rush, J. D. (2009). The meta-analysis of clinical judgment project: Effects of experience on judgment accuracy. The Counseling Psychologist, 37(3), 350-399.

Spohr, D. (2017). Fake news and ideological polarization: Filter bubbles and selective exposure on social media. Business Information Review, 34(3), 150-160.

Stagnaro, MN, Pennycook, G., & Rand, DG (2018) Performance on the Cognitive Reflection Test is stable across time. Judgment and Decision Making, 13, 260-267.

Stanovich, K. E. (2021). The bias that divides us: The science and politics of myside thinking. MIT Press.

Stanovich, K. E., & West, R. F. (2007). Natural myside bias is independent of cognitive ability. Thinking & Reasoning, 13(3), 225-247.

Stanovich, K. E., & West, R. F. (2008). On the relative independence of thinking biases and cognitive ability. Journal of personality and social psychology, 94(4), 672.

Stanovich, K. E., West, R. F., & Toplak, M. E. (2013). Myside bias, rational thinking, and intelligence. Current Directions in Psychological Science, 22(4), 259–264. https://doi.org/10.1177/0963721413480174

Stieglitz, S., Brachten, F., Ross, B., & Jung, A. K. (2017). Do social bots dream of electric sheep? A categorisation of social media bot accounts. arXiv preprint arXiv:1710.04044.

Stocking, G., & Sumida, N. (2018). Social media bots draw public's attention and concern. Pew Research Center's Journalism Project, 15.

Strickland, A. A., Taber, C. S., & Lodge, M. (2011). Motivated reasoning and public opinion. Journal of health politics, policy and law, 36(6), 935-944.

Stroud, N. J. (2008). Media use and political predispositions: Revisiting the concept of selective exposure. Political Behavior, 30(3), 341-366.

Sutton, R. T., Pincock, D., Baumgart, D. C., Sadowski, D. C., Fedorak, R. N., & Kroeker, K. I. (2020). An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ digital medicine, 3(1), 1-10.

Taber, C. S., & Lodge, M. (2006). Motivated skepticism in the evaluation of political beliefs. American journal of political science, 50(3), 755-769.

Taber, C. S., Cann, D., & Kucsova, S. (2009). The motivated processing of political arguments. Political Behavior, 31(2), 137-155.

Tambe, P., Cappelli, P., & Yakubovich, V. (2019). Artificial intelligence in human resources management: Challenges and a path forward. California Management Review, 61(4), 15-42.

Tappin, B. M., Pennycook, G., & Rand, D. G. (2020). Rethinking the link between cognitive sophistication and politically motivated reasoning. Journal of Experimental Psychology: General.

Thaler, R. H., & Sunstein, C. R. (2008) Nudge: Improving decisions about health, wealth, and happiness. Yale University Press, New Haven, CT.

Thaler, R. H., & Ganser, L. J. (2015). Misbehaving: The making of behavioral economics.

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

Timberg, C., & Dwoskin, E. (2018). Twitter is sweeping out fake accounts like never before, putting user growth at risk. Washington Post. Retrieved November 3, 2020. https://www.washingtonpost.com/technology/2018/07/06/twitter-is-sweeping-out-fake-accounts-like-never-before-putting-user-growth-risk/

Toplak, M. E., & Stanovich, K. E. (2003). Associations between myside bias on an informal reasoning task and amount of post-secondary education. Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition, 17(7), 851-860.

Toplak, M. E., West, R. F., & Stanovich, K. E. (2011). The Cognitive Reflection Test as a predictor of performance on heuristics-and-biases tasks. Memory & cognition, 39(7), 1275-1289.

Tramèr, F., Kurakin, A., Papernot, N., Goodfellow, I., Boneh, D., & McDaniel, P. (2017). Ensemble adversarial training: Attacks and defenses. arXiv preprint arXiv:1705.07204.

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. Journal of Risk and uncertainty, 5(4), 297-323.

Twitter. (2019). Elections integrity data archive. https : / / about.twitter.com /en us/values/elections-integrity.html#us-elections. (Accessed on 03/30/2019).

Twitter. [@TwitterSupport]. (2021, September 9). What's a bot and what's not? We're making it easier to identify #GoodBots and their automated Tweets with new labels. [Tweet]. Twitter. https://twitter.com/TwitterSupport/status/1436073604173770754

Tyagi, A., Babcock, M., Carley, K. M., & Sicker, D. C. (2020, October). Polarizing tweets on climate change. In International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in

Modeling and Simulation (pp. 107-117). Springer, Cham.

Uyheng, J & Carley, K.M. (2020). Bots and online hate during the COVID-19 pandemic: Case studies in the United States and the Philippines. Journal of Computational Social Science, 3(2): 445-468.

Uyheng, J & Carley, K.M. (2020). "Bot Impacts on Public Sentiment and Community Structures: Comparative Analysis of Three Elections in the Asia-Pacific," In Proceedings of the International Conference SBP-BRiMS 2020, Halil Bisgin, Ayaz Hyder, Chris Dancy, and Robert Thomson (Eds.) Washington DC, October 2020, Springer.

Uyheng, J & Carley, K.M. (2021). Characterizing network dynamics of online hate communities around the COVID-19 pandemic. Applied Network Science, 6(20).

Uyheng, J., Ng, L. H. X., & Carley, K. M. (2021). Active, aggressive, but to little avail: characterizing bot activity during the 2020 Singaporean elections. Computational and Mathematical Organization Theory, 27(3), 324-342.

Varol, O., Ferrara, E., Davis, C. A., Menczer, F., & Flammini, A. (2017). Online Human-Bot Interactions: Detection, Estimation, and Characterization. Proceedings of the Eleventh International AAAI Conference on Web and Social Media.

Veale, T., & Cook, M. (2018). Twitterbots: Making Machines that make meaning. MIT Press.

Venkatesh, V., Thong, J. Y., & Xu, X. (2016). Unified theory of acceptance and use of technology: A synthesis and the road ahead. Journal of the association for Information Systems, 17(5), 328-376.

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. Science, 359(6380), 1146-1151.

Wald, R., Khoshgoftaar, T.M., Napolitano, A., & Sumner, C. (2013). "Predicting susceptibility to social bots on twitter." In 2013 IEEE 14th International Conference on Information Reuse & Integration (IRI), pp. 6-13. IEEE.

Wang, P., Angarita, R., & Renna, I. (2018). Is this the era of misinformation yet: combining social bots and fake news to deceive the masses. In Companion Proceedings of the The Web Conference 2018 (pp. 1557-1561).

Warm, J. S., Parasuraman, R., & Matthews, G. (2008). Vigilance requires hard mental work and is stressful. Human factors, 50(3), 433-441.

Warren-West, L. S., & Jackson, R. C. (2020). Seeing the bigger picture: susceptibility to, and detection of, deception. Journal of Sport and Exercise Psychology, 42(6), 463-471.

Weiss, D. J., & Shanteau, J. (2003). Empirical Assessment of Expertise. Human Factors: The Journal of the Human Factors and Ergonomics Society, 45(1), 104–116. https://doi.org/10.1518/hfes.45.1.104.27233

Westerwick, A., Johnson, B. K., & Knobloch-Westerwick, S. (2017). Confirmation biases in selective exposure to political online information: Source bias vs. content bias. Communication Monographs, 84(3), 343–364. https://doi.org/10.1080/03637751.2016.1272761

Williams, M. (2001). In whom we trust: Group membership as an affective context for trust development. Academy of management review, 26(3), 377-396.

Witteman, C. L., & Tollenaar, M. S. (2012). Remembering and diagnosing clients: Does experience matter? Memory, 20(3), 266-276.

Wiyatno, R. R., Xu, A., Dia, O., & de Berker, A. (2019). Adversarial examples in modern machine learning: A review. arXiv preprint arXiv:1911.05268.

Wolfe, J. M., Brunelli, D. N., Rubinstein, J., & Horowitz, T. S. (2013). Prevalence effects in newly trained airport checkpoint screeners: Trained observers miss rare targets, too. Journal of Vision, 13(3), 1–9. doi:10.1167/13.3.33

Wu, L., Morstatter, F., Carley, K. M., & Liu, H. (2019). Misinformation in social media: definition, manipulation, and detection. ACM SIGKDD Explorations Newsletter, 21(2), 80-90.

Yang, K. C., Ferrara, E., & Menczer, F. (2022). Botometer 101: Social bot practicum for computational social scientists. arXiv preprint arXiv:2201.01608.

Yokoi, R., Eguchi, Y., Fujita, T., & Nakayachi, K. (2021). Artificial intelligence is trusted less than a doctor in medical treatment decisions: Influence of perceived care and value similarity. International Journal of Human–Computer Interaction, 37(10), 981 990.

Zhang, Y., Lukito, J., Su, M. H., Suk, J., Xia, Y., Kim, S. J., ... & Wells, C. (2021). Assembling the networks and audiences of disinformation: How successful Russian IRA Twitter accounts built their followings, 2015–2017. Journal of Communication, 71(2), 305-331.

Zhang, Y., Shah, D., Pevehouse, J., & Valenzuela, S. (2022). Reactive and Asymmetric Communication Flows: Social Media Discourse and Partisan News Framing in the Wake of Mass Shootings. The International Journal of Press/Politics, 19401612211072793.