Duped by Bots: Why Some are Better than Others at Detecting Fake Social Media Personas

Ryan Kenny, Baruch Fischhoff, Alex Davis and **Kathleen M. Carley**, Carnegie Mellon University, Pittsburgh, PA, USA, **Casey Canfield**, Missouri University of Science and Technology, Rolla, MI, USA

Objective: We examine individuals' ability to detect social bots among Twitter personas, along with participant and persona features associated with that ability.

Background: Social media users need to distinguish bots from human users. We develop and demonstrate a methodology for assessing those abilities, with a simulated social media task.

Method: We analyze performance from a signal detection theory perspective, using a task that 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. We estimated the probability of participants indicating that a persona was a bot with a generalized linear mixed-effects model using participant characteristics (social media experience, analytical reasoning, and political views) and stimulus characteristics (bot indicator score and political tone) as regressors.

Results: 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 identity.

Conclusions: Individuals have limited ability to detect social bots, with greater aversion to mistaking bots for humans than vice versa. Greater social media experience and myside bias appeared to reduce performance, as did less analytical reasoning ability.

Application: These patterns suggest the need for interventions, especially when users feel most familiar with social media.

Keywords: signal detection theory, social bots, social media, analytical reasoning, myside bias

Address correspondence to Ryan Kenny, Carnegie Mellon University, 5215 Wean Hall, Pittsburgh, PA 15213-3815, USA; e-mail: ryankenn@andrew.cmu.edu

HUMAN FACTORS
2024, Vol. 66(1) 88–102
DOI:10.1177/00187208211072642
Article reuse guidelines: sagepub.com/journals-permissions

INTRODUCTION

Social bots are artificial agents that infiltrate social media (Cresci, 2020). Although most social bots are relatively harmless and execute mundane advertising tasks (Appel et al., 2020), some are designed to manipulate social media discourse by disseminating misinformation, encouraging false beliefs, or discrediting valid sources of information (Ferrara et al., 2016; Huang, & Carley, 2020; Pacheco et al., 2020; Wu et al., 2019).

Social bots can accomplish these goals indirectly, by creating false impressions of support for a persona or position, or directly by promulgating lies and half-truths in coordinated campaigns. Recent analyses have found that social bots play a disproportionate role in proliferating low credibility information in social media, potentially influencing socio-political events (Caldarelli et al., 2020; Shao et al., 2018; Shorey & Howard, 2016) and spreading hate (Uyheng & Carley, 2020b, 2021). Their impact is amplified by bad actors' ability to scale deployment at little cost, while retaining anonymity and avoiding accountability (Cresci, 2020; Veale & Cook, 2018).

Bot activity is notably widespread on Twitter, a social media microblogging and social networking service employing short (280 character or fewer) messages, called tweets (Chu et al., 2010). In 2018, Twitter disclosed that it had deleted approximately 70 million fake accounts (Timberg & Dwoskin, 2018). Analysts estimate that 9%–15% of active Twitter accounts are non-human bots (Varol et al., 2017). Twitter activity involving social bots targeting socio-political activity has been estimated as high as 25%–30% (Huang, 2020; Uyheng & Carley, 2020a, 2020b).

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 degrees of human oversight and control (Stieglitz et al., 2017). Twitter users seeking large audiences of followers (*influencers*) can employ social bots to simulate engagement and flood social networks with content, while drowning out other voices (Cook et al., 2014; Jansen et al., 2009; Lee et al., 2010; Riquelme, & González-Cantergiani, 2016).

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 et al., 2017).

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, 2018a) and Botometer (Davis et al., 2016). The two systems were developed independently and trained with different data. Both produce a probability for a persona being a social bot. We combined them to produce a *bot indicator score* for each persona.

Botometer uses six categories of features as evidence for its bot probability score (Davis et al, 2016): *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 originally trained on a data set generated using honeypots to attract social bots (Lee et al., 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, 2018a). The bot indicator scores used here were based on predictions using lower tier 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, 2018a, 2018b), 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%. As there were relatively few stimuli with high Bot-hunter scores (>85%), we relaxed the criterion for eight stimuli in that range. Botometer still gave high probabilities for those personas being bots, just lower ones than with Bot-hunter. For these stimuli, we used the Bot-Hunter probability as the bot indicator score. 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, and correct rejections). If so, and they believe that bot and human stimuli are equally common, they should respond "bot" and "human" equally often (Canfield et al., 2016). Choosing "bot" more often would suggest an aversion to misidentifying one of them; similarly, with 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 (Endsley, 2018; Landy, 2018; Weiss & Shanteau, 2003). We did not expect social media experience to affect participants' decision criterion.

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 found to perform better on discrimination and judgment tasks similar to the present one (Bar-Hillel et al., 2019; Campitelli & Labollita, 2010; Toplak et al., 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' criterion.

Myside Bias: Social media users tend to follow and engage with sources that agree with them (Bakshy et al., 2015; Flaxman et al., 2016; Stroud, 2008). 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 (Drummond & Fischhoff, 2019; Kahn & Davies, 2011; Stanovich et al., 2013; Stanovich & West, 2008; Toplak & Stanovich, 2003). 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 (Barberá et al., 2015; Westerwick et al., 2017).

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 answer more attention checks correctly to demonstrate greater sensitivity (Dewitt et al., 2015; Downs et al., 2010; Matthews et al., 2010). 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 et al., 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, and follower count). Although all personas were from real-world accounts active when their bot indicator scores were obtained (see above), participants were not told this explicitly. Twenty-five trials were "bot" personas; 25 were "human," as defined by bot indicator scores above or below 50%, respectively. 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 post-hoc analyses. Full instructions appear in the Supplementary Materials (SM).

After completing the judgment task, participants completed a demographic survey and individual difference measures of (a) social media experience (Hou, 2017) (see Supplementary Materials); (b) political views, on a 5-point scale ranging from "liberal" to "conservative"; and (c) cognitive reflection tendency, using the original three-item CRT (Frederick, 2005). A *political difference* score was created as the absolute difference between 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 characteristics (task engagement, social media experience, cognitive reflection score, and political difference), and their interactions to predict the sensitivity and criterion of participants' judgments of whether each stimulus was a bot or human persona. Post-hoc analysis examined relationships between participants' political views and their performance.

Traditional SDT analyses estimates 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),

Figure 1. Example of Twitter Persona Profile. *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.

assuming a multivariate normal distribution (DeCarlo, 1998). Equations (1) and (2) show the planned and post-hoc models, following Farewell, Long, Tom, Yiu and Su (2017):

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

$$\begin{aligned} &P(\text{`bot'}Bi_{\text{i,}}TE_{\text{i,}}TO_{\text{i,}}SME_{\text{i,}}CRT_{\text{i,}}PV_{\text{i,}}PD_{\text{i,}}) = \\ &\Phi(\beta_0 + \beta_1 Bi_i + \beta_2 TE_i + \beta_3 TO_i + \beta_4 SME_i \\ &+\beta_5 CRT_i + \beta_6 PV_i + \beta_7 PD_i + \beta_8 TE_i{*}Bi_i \\ &+\beta_9 TO_i{*}Bi_i + \beta_{10} SME_i{*}Bi_i + \beta_{11} CRT_i{*}Bi_i \\ &+\beta_{12} PV_i{*}Bi_i + \beta_{13} PD_i{*}Bi_i \\ &+\beta_{14} PD_i{*}CRT_i{*}Bi_i + \beta_{15} PV_i{*}CRT_i{*}Bi_i \\ &+\beta_{16} PD_i{*}PV_i{*}Bi_i + \beta_{17} PV_i{*}PD_i{*}CRT_i{*}Bi_i) \end{aligned} \tag{2}$$

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 Θ. Equation (2) shows how the probability of responding "bot" is computed, applying the probit link function to the set of conditioned predictor variables. 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 et al., 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., stronger 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.

As mentioned, 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, respectively. 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 results of both the planned and the post-hoc analyses, with the latter adding respondents' political values (PV).

RESULTS

Sample Demographics

The sample included 113 participants, 73 men and 40 women, whose age ranged from 18 to 72 years old (mean = 36 and 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, seven employed part time, one unemployed but looking, two retired, eight were self-employed, and one 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 and 60K.

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 coefficient for Task Engagement (TE) 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).

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, versus "human" = 0, as a function of task characteristics (bot indicator, task order), individual characteristics (task engagement, social media experience, cognitive reflection test, and 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 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.

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 both 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 did conservatives.

The interaction between bot indicator and political differences (Bi x PD) in both models 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.

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 CRT score increased, so did sensitivity to Bi for liberals, but not for conservatives.

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.

DISCUSSION

This study assesses human performance in detecting bots among Twitter personas. Participants judged whether each of 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

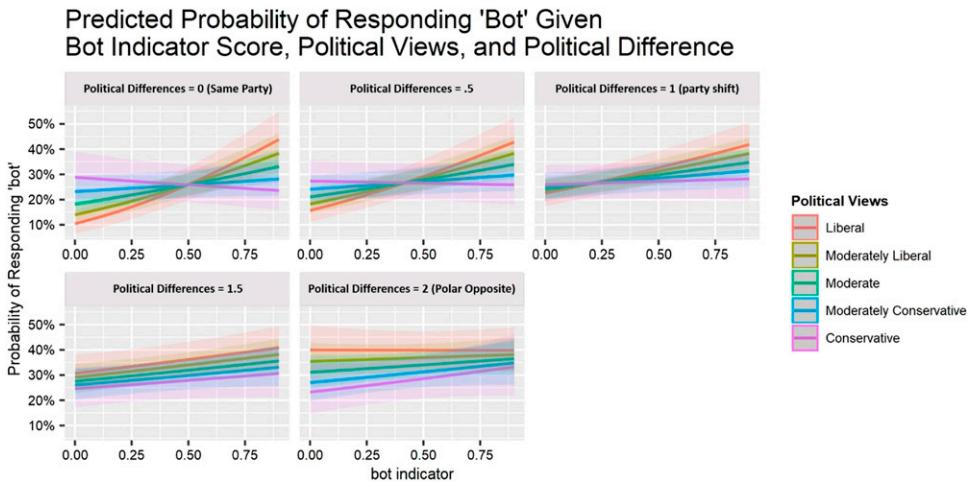

Figure 2. Predicted probability of Responding "bot" given bot indicator score, political views, and political difference.

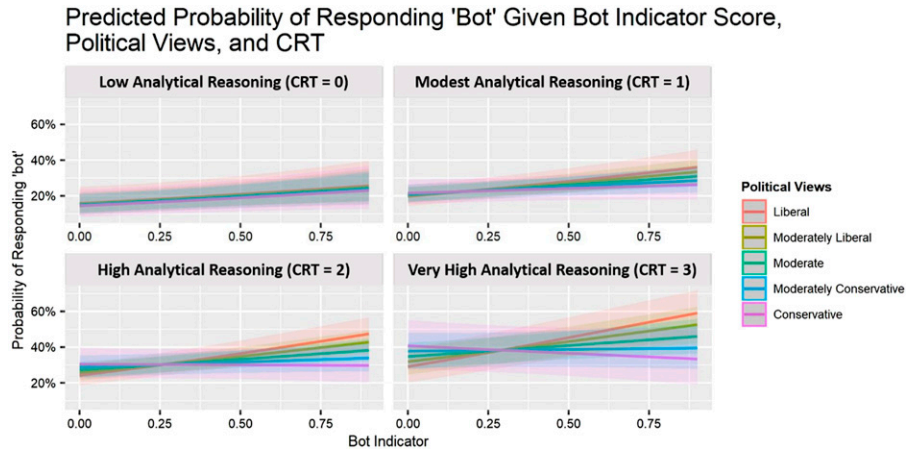

Figure 3. Predicted probability of responding "bot" given bot indicator score, political views, and CRT.

score, this 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 et al., 2004; Bond, 2008; Spengler et al., 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 et al., 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.

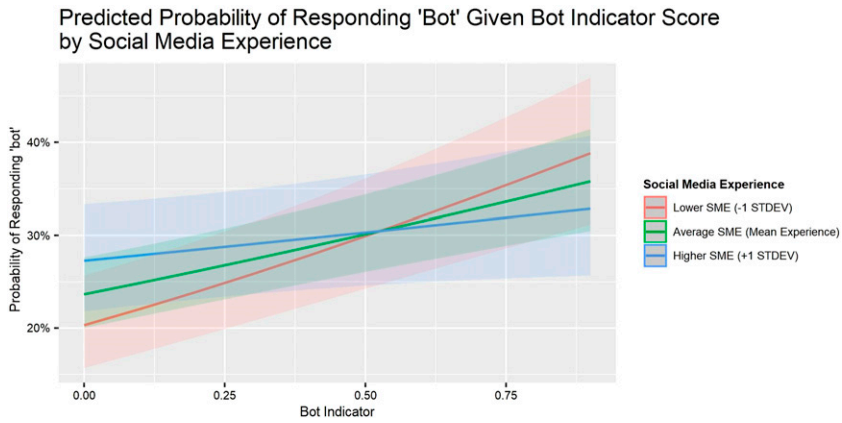

Figure 4. Predicted probability of responding "bot" given bot indicator score by social media experience.

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 that of the persona. 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 also 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, when judging politically similar personas, liberals were more sensitive to bot indicator (Bi) than were conservatives. 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 et al. 2019). These results are consistent with that distrust, revealing themselves somewhat differently with liberals than conservatives, suggesting either differences in reasoning styles (Deppe et al., 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 (Bronstein et al., 2019; Pennycook & Rand, 2019; Ross et al., 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, 2007K. E. Stanovich & West, 2007; Strickland et al., 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 which may be bots (Choi et al., 2020; Sasahara et al., 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 1 year after their Twitter profiles were initially assessed. 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 (Aiello et al., 2012; Wald et al., 2013), 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 normal 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 et al., 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 persona, attempting to manipulate users' commercial or 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 or bot indicator scores. Evaluating such intervention is an urgent question for protecting people from social media manipulation.

ACKNOWLEDGMENTS

Funding for this research was provided by the U.S. Army, Advanced Strategic Planning and Policy Program, Goodpaster Fellowship, and by the Swedish Foundation for Humanities and Social Science. Additionally, we would like to thank Dave Beskow for access to Bot-hunter and his early advice on this endeavor, and Megan Kenny, as independent coder. The views expressed are those of the authors.

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 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.

ORCID iD

Casey Canfield https://orcid.org/0000-0001-5325-3798

REFERENCES

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, Dublin, Ireland, 4–8 June *2012*, (Vol. 6, No. 1).

Allen, R., Mcgeorge, P., Pearson, D., & Milne, A. B. (2004). Attention and expertise in multiple target tracking. *Applied Cognitive Psychology*, *18*(3), 337–347. https://doi.org/10.1002/acp.975

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. https://doi.org/10.1007/s11747-019-00695-1

Bakshy, E., Messing, S., & Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on facebook. *Science*, *348*(6239), 1130–1132. https://doi.org/10.1126/science.aaa1160.

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. https://doi.org/10.1177/0956797615594620

Bar-Hillel, M., Noah, T., & Frederick, S. (2019). Solving stumpers, CRT and CRAT: Are the abilities related? *Judgment and Decision Making*, *14*(5), 620-623.

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. In *Conference: SBP-BRiMS: International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*, Washington, DC, 10-13th July 2018.

Beskow, D., & Carley, K. (2018b). Introducing bothunter: A tiered approach to detection and characterizing automated activity on twitter. In H Bisgin, A Hyder, C Dancy, & R Thomson (Eds.), *International conference on social computing, behavioral-cultural modeling and prediction and behavior representation in modeling and simulation*. Springer.

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. https://doi.org/10.1080/00140138108924833

Bond, G. D. (2008). Deception detection expertise. *Law and Human Behavior*, *32*(4), 339–351. https://doi.org/10.1007/s10979-007-9110-z

Botometer (2021). CNetS. https://cnets.indiana.edu/blog/tag/botometer/

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. https://doi.org/10.1016/j.jarmac.2018.09.005

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), 81. https://doi.org/10.1038/s42005-020-0340-4.

Campitelli, G., & Labollita, M. (2010). Correlations of cognitive reflection with judgments and choices. *Judgment and Decision Making*, *5*(3), 10.

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

Canfield, C. I., Fischhoff, B., & Davis, A. (2016). Quantifying phishing susceptibility for detection and behavior decisions. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *58*(8), 1158–1172. https://doi.org/10.1177/0018720816665025

Choi, D., Chun, S., Oh, H., Han, J., & Kwon, T. T. (2020). Rumor propagation is amplified by echo chambers in social media. *Scientific Reports*, *10*(1), 310. https://doi.org/10.1038/s41598-019-57272-3

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, San Juan, PR, USA, 03–07 December, 2018 (pp. 21–30).

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. https://www.jstor.org/stable/26487011

Cresci, S. (2020). A decade of social bot detection. *Communications of the ACM*, *63*(10), 72–83. https://doi.org/10.1145/3409116

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. https://doi.org/10.1371/journal.pone.0057410

Davis, C. A., Varol, O., Ferrara, E., Flammini, A., & Menczer, F. (2016). BotOrNot: A system to evaluate social bots. In Proceedings of the 25th international conference companion on world wide web, Montréal, Canada, 11–15, May, 2016, WWW '16 Companion, 273–274 https://doi.org/10.1145/2872518.2889302

DeCarlo, L. T. (1998). Signal detection theory and generalized linear models. *Psychological Methods*, *3*(2), 186–205. https://doi.org/10.1037/1082-989x.3.2.186

Deppe, K. D., Gonzalez, Neiman, J., Pahlke, J., Smith, K., & Hibbing, J. R. (2015). Reflective liberals and intuitive conservatives: A look at the cognitive reflection test and ideology. *Judgment and Decision Making*, *10*(4), 314-331.

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. https://doi.org/10.1088/1748-9326/10/12/124009

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*. ACM (pp. 2399–2402).

Drummond, C., & Fischhoff, B. (2017). Individuals with greater science literacy and education have more polarized beliefs on controversial science topics. *Proceedings of the National Academy of Sciences*, *25*(4), 9587-9592. https://doi.org/10.1073/pnas.1704882114.

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

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.. Cambridge University Press, pp. 714–742. https://doi.org/10.1017/9781316480748.037

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*. Cambridge University Press (pp. 745–769). https://doi.org/10.1017/9781316480748.038

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. https://doi.org/10.1146/annurev-statistics-060116-054131.

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. https://doi.org/10.1093/poq/nfw006

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, *19*(4), 25–42. https://doi.org/10.1257/089533005775196732

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. *Parasites & Vectors*, *1*, 9-3. https://cran.r-project.org/web/packages/arm/index.html.

Gramlich, J. (2020). *Democrats, Republicans each expect made-up news to target their own party more than the other in 2020*. 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*. Wiley.

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

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

Huang, B., (2020). "*Learning user latent attributes on social media, "Ph.D Thesis*. School of Computer Science, Institute of Software Research, Carnegie Mellon University.

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*(1), 129–146. https://doi.org/10.1146/annurev-polisci-051117-073034

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. https://doi.org/10.1002/asi.21149

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. https://doi.org/10.1177/1368430210374609

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.. Cambridge University Press, pp. 151–164. https://doi.org/10.1017/9781316480748.010

Langley, P. (1985). Learning to search: from weak methods to domain-specific heuristics*. *Cognitive Science*, *9*(2), 217–260. https://doi.org/10.1207/s15516709cog0902_2

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, North Carolina, USA, 26–30 April 2010 (pp. 1137–1138) for detailed perspectives on influence in social media environments. https://doi.org/10.1145/1772690.1772842

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, Barcelona, Catalonia, 17–21 July 2011, *(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. *Sociological Forum*, *35*(1), 996–1018). https://doi.org/10.1111/socf.12603

Loepp, E., & Kelly, J. T. (2020). Distinction without a difference? An assessment of MTurk worker types. *Research & Politics*, *7*(1), 2053168019901185. https://doi.org/10.1177/2053168019901185

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. https://doi.org/10.3758/s13428-011-0124-6

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*.

Springer (pp. 205–230). https://doi.org/10.1007/978-1-4419-1210-7_13

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

Pacheco, D., Hui, P.-M., Torres-Lugo, C., Truong, B. T., Flammini, A., & Menczer, F. (2020). *Uncovering coordinated networks on social media*. ArXiv:2001.05658 [Physics] http://arxiv.org/abs/2001.05658

Parasuraman, R., & Davies, D. R. (1977). A taxonomic analysis of vigilance performance. In *Vigilance*. Springer (pp. 559–574). https://doi.org/10.1007/978-1-4684-2529-1_26

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. https://doi.org/10.1016/j.cognition.2018.06.011.

Rikers, R. M., Schmidt, H. G., & Boshuizen, H. P. (2000). Knowledge encapsulation and the intermediate effect. *Contemporary Educational Psychology*, *25*(2), 150–166. https://doi.org/10.1006/ceps.1998.1000

Riquelme, F., & González-Cantergiani, P. (2016). Measuring user influence on twitter: A survey. *Information Processing & Management*, *52*(5), 949–975. https://doi.org/10.1016/j.ipm.2016.04.003

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.

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

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

Shorey, S., & Howard, P. N. (2016). Automation, big data, and politics: A research review. *International Journal of Communication*, *10*(1), 5032–5055.

Spengler, P. M., White, M. J., Ægisdóttir, S., Maugherman, A. S., Anderson, L. A., Cook, R. S., Nichols, C. N., Lampropoulos, G. K., Walker, B. S., Cohen, G. R., & Rush, J. D. (2009). The meta-analysis of clinical judgment project. *The Counseling Psychologist*, *37*(3), 350–399. https://doi.org/10.1177/0011000006295149

Stanovich, K. E., & West, R. F. (2007). Natural myside bias is independent of cognitive ability. *Thinking & Reasoning*, *13*(3), 225–247. https://doi.org/10.1080/13546780600780796

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), 672–695. https://doi.org/10.1037/0022-3514.94.4.672

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. https://doi.org/10.1215/03616878-1460524

Stroud, N. J. (2008). Media use and political predispositions: Revisiting the concept of selective exposure. *Political Behavior*, *30*(3), 341–366. https://doi.org/10.1007/s11109-007-9050-9

Timberg, C., & Dwoskin, E. (2018). Twitter is sweeping out fake accounts like never before, putting user growth at risk. *Washington Post*. 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*, *17*(7), 851–860. https://doi.org/10.1002/acp.915.

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. https://doi.org/10.3758/s13421-011-0104-1

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

Uyheng, J., & Carley, K. M. (2020a). 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*, Washington, DC, USA, October *18-21*, 2020. https://doi.org/10.1007/978-3-030-61255-9_2

Uyheng, J., & Carley, K. M. (2020b). 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), 1-24. https://doi.org/10.1007/s42001-020-00087-4.

Uyheng, J., & Carley, K. M. (2021). Characterizing network dynamics of online hate communities around the COVID-19 pandemic. *Applied Network Science*, *6*(1), 20. https://doi.org/10.1007/s41109-021-00362-x

Varol, O., Ferrara, E., Davis, C. A., Menczer, F., & Flammini, A. (2017). Online human-bot interactions: Detection, estimation, and characterization. In *Proceedings of the Eleventh International AAAI Conference on Web and Social Media*. arXiv:1703.03107.

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

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)*, San Francisco, 14–16 August 2013 (pp. 6–13). IEEE. https://doi.org/10.1109/iri.2013.6642447

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

Witteman, C. L., & Tollenaar, M. S. (2012). Remembering and diagnosing clients: Does experience matter? *Memory*, *20*(3), 266–276. https://doi.org/10.1080/09658211.2012.654799

Wu, L., Morstatter, F., Carley, K. M., & Liu, H. (2019). Misinformation in social media. *ACM SIGKDD Explorations Newsletter*, *21*(2), 80–90. https://doi.org/10.1145/3373464.3373475

Ryan Kenny. Ryan Kenny is a Lieutenant Colonel in the United States Army, serving in the Signal Corps. He received a BA in cognitive psychology from the University of Notre Dame, in 2003, an MA in national security and strategic studies from the U.S. Naval War College, Newport, RI in 2015, and is pursuing his PhD in Engineering and Public Policy at Carnegie Mellon University, Pittsburgh, PA. His research interests include Human–machine Systems, Artificial Intelligence, and Behavioral Decision-Making.

Baruch Fischhoff. Baruch Fischhoff is Professor in the Department of Engineering and Public Policy and Institute for Politics and Strategy, Carnegie Mellon University. He studies decision-making, with a focus on empowering people to participate actively in public and private decisions. He went to the Detroit Public Schools, Wayne State University (mathematics, psychology), and the Hebrew University of Jerusalem (psychology). He is an elected member of the National Academy of Sciences and of the National Academy of Medicine. His books include *Acceptable Risk, Risk: A Very Short Introduction*, *Risk Communications: The Mental Models Approach*, and *Counting Civilian Casualties*.

Alex Davis. Alex Davis is Associate Professor in the Department of Engineering and Public Policy, Carnegie Mellon University. He studies decision-making with a focus on statistical modeling. He is a graduate of Northern Arizona University (B.S in psychology) and Carnegie Mellon University (PhD in behavioral decision-making). His research includes using statistical models to improve risk communication during pregnancy, statistical and behavioral models of individual and group preference, and the integration of human decision-making with artificial intelligence.

Kathleen Carley. Kathleen M. Carley is Professor of Societal Computing, School of Computer Science's Institute for Software Research, Carnegie Mellon University; the Director of the Center for Computational Analysis of Social and Organizational Systems (CASOS); and Director of the Center for Informed Democracy and Social Cybersecurity (IDeaS), studying disinformation, hate speech and extremism online. She has a Ph.D. from Harvard, and two S.B.s from the Massachusetts Institute of Technology. Her research interests include applying computational social science, cognitive science, organization science, dynamic network analysis, social network analysis, machine learning, data analytics, and text analytics to complex social and organizational problems.

Casey Canfield. Casey Canfield is an Assistant Professor in Engineering Management & Systems Engineering at Missouri University of Science & Technology. She has a B.S. in Engineering: Systems from Olin College of Engineering and a Ph.D. in Engineering and Public Policy from Carnegie Mellon University. Her research focuses on quantifying the human part of complex systems to improve decision-making at individual and organizational levels in the context of energy, rural broadband, governance, and healthcare.