01 Jan 2023

Improving Social Bot Detection Through Aid And Training

Ryan Kenny

Baruch Fischhoff

Alex Davis

Casey I. Canfield
Missouri University of Science and Technology, canfieldci@mst.edu

Recommended Citation

R. Kenny et al., "Improving Social Bot Detection Through Aid And Training," *Human Factors*, SAGE Publications; Human Factors and Ergonomics Society, Jan 2023.
The definitive version is available at https://doi.org/10.1177/00187208231210145

Extended Multi-Phase Study

Human Factors
2023, Vol. 0(0) 1–22

DOI: 10.1177/00187208231210145
journals.sagepub.com/home/hfs


Improving Social Bot Detection Through Aid and Training

Ryan Kenny[1], Baruch Fischhoff[2], Alex Davis[2], and Casey Canfield[3]

Abstract

Objective: We test the effects of three aids on individuals' ability to detect social bots among Twitter personas: a bot indicator score, a training video, and a warning.

Background: Detecting social bots can prevent online deception. We use a simulated social media task to evaluate three aids.

Method: Lay participants judged whether each of 60 Twitter personas was a human or social bot in a simulated online environment, using agreement between three machine learning algorithms to estimate the probability of each persona being a bot. Experiment 1 compared a control group and two intervention groups, one provided a bot indicator score for each tweet; the other provided a warning about social bots. Experiment 2 compared a control group and two intervention groups, one receiving the bot indicator scores and the other a training video, focused on heuristics for identifying social bots.

Results: The bot indicator score intervention improved predictive performance and reduced overconfidence in both experiments. The training video was also effective, although somewhat less so. The warning had no effect. Participants rarely reported willingness to share content for a persona that they labeled as a bot, even when they agreed with it.

Conclusions: Informative interventions improved social bot detection; warning alone did not.

Application: We offer an experimental testbed and methodology that can be used to evaluate and refine interventions designed to reduce vulnerability to social bots. We show the value of two interventions that could be applied in many settings.

Keywords

signal detection theory, social bots, social media, decision support, training

Introduction

The rise of bots continues—particularly in online environments. When faced with uncertainty about the legitimacy of online interactions, social media users face a Turing Test, deciding whether they believe that online personas are humans or social bots. People often fail this test. As a result, social bots go unnoticed, misleading those who place unwarranted trust in them (Cresci, 2020; Stieglitz et al., 2017).

[1]United States Army, Fayetteville, NC, USA
[2]Carnegie Mellon University, Pittsburgh, PA, USA
[3]Missouri University of Science and Technology, Rolla, MO, USA

Received: October 10, 2022; accepted: October 9, 2023

Corresponding Author:
Baruch Fischhoff, Department of Engineering and Public Policy, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA 15213-3815, USA; e-mail: baruch@cmu.edu

A recent article in the Wall Street Journal noted the increasing role of social bots in directing information within social media spaces, particularly Twitter (Cohen, 2023; also, Harari, 2023; Satariano & Mozur, 2023; Thompson, 2023).

Social bots flood social networks with content, attempting to drown out other voices (Cook et al., 2014; Jansen et al., 2009; Lee et al., 2010; Riquelme & González-Cantergiani, 2016). In online environments, where attention is commoditized, limited attention in social networks (Hodas & Lerman, 2012; Lerman et al., 2013) can incentivize the use of automated tools. Most bots execute innocuous advertising tasks (Appel et al., 2020). Others, though, pursue socially harmful ends, such as spreading misinformation and encouraging false beliefs (Cresci, 2020; Ferrara et al., 2016; Huang & Carley, 2020; Pacheco et al., 2020; Pennycook & Rand, 2019; Shao et al., 2018; Wu et al., 2019).

Twitter has had notably nefarious social bot activity (Chu et al., 2010; Huang, 2020; Timberg & Dwoskin, 2018; Uyheng & Carley, 2020; Varol et al., 2017). Online armies of fully and semi-automated social bots seek to inflate follower counts, generate message "likes," and induce other users to share, or "retweet," content (Stieglitz et al., 2017).

The first step in degrading social bots' impact is detecting them. The primary cues for detection are in the persona accompanying each tweet. Observers must examine those cues and decide whether to treat a persona as a human or a bot, considering the costs and benefits of correct and incorrect decisions. Signal detection theory (SDT) formalizes such tasks in terms of two components: (a) *sensitivity*, reflecting how well an observer can distinguish the two possible states; and (b) *criterion*, reflecting the observer's preferences for making correct judgments and avoiding incorrect ones (Green & Swets, 1966; Macmillan & Creelman, 2004; McNicol, 2005; Stanislaw & Todorov, 1999). These factors apply whether human observers or machine learning algorithms are the detectors (Batailler et al., 2022).

Signal detection theory has been used to assess the factors influencing observers' sensitivity and criteria when assessing interpersonal deception (Bond, 2008; Bond & DePaulo, 2006; Hauch et al., 2016; Warren-West & Jackson, 2020). In an earlier study, we extended SDT to study social bot detection (Kenny et al., In press). Using a platform simulating online Twitter use, we found low sensitivity, with a criterion biased toward saying "human," suggesting that users were more averse to dismissing a human as a bot than vice versa. There was considerable variability, with greater sensitivity among individuals who (1) reported less social media experience, (2) were evaluating personas with opposing political beliefs, and (3) scored higher on a test of critical reasoning ability.

Our extension of SDT addressed an issue common in online environments, the lack of a ground truth for evaluating users' performance. That is, no one but a persona's creator typically knows whether it represents a human (Meaker, 2022). As a way to approximate ground truth, we used the convergence of three bot detection algorithms that use different computational procedures and were trained on different data sets. We hoped to reduce shared errors, while recognizing that black-box algorithms may still hide them.

Building on our previous study, we use the experimental platform to examine the efficacy of three interventions for improving social bot detection: a repetitive warning about social bots, accompanying each persona; a bot indicator score for each persona, reflecting machine learning algorithm predictions; and an introductory training video, offering heuristics for bot detection. Our platform uses Twitter persona profiles with persona cues that users could readily find on actual systems. Figure 1 illustrates such a persona.

Experiment 1

Computer scientists have applied machine learning (ML) methods to develop social bot detection tools, including Botometer (Botometer, 2021; Davis et al., 2016; Yang et al., 2022), Bot Hunter (Beskow & Carley, 2018a, 2018b), and Bot Sight (Kats, 2022). Our previous study (Kenny et al., In press) used agreement among these three algorithms to identify personas with a range of probabilities of being a bot. Our current study uses those algorithms to create interventions designed to improve performance.

Machine learning algorithms claim greater than 90% accuracy in detecting social bots. However, they use cues unavailable to ordinary

Figure 1. Example of Twitter persona profile. *Note.* The arrows indicate Twitter persona features. 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 persona owner 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.

Twitter users, including both persona-defined information (i.e., tweet text and account details such as persona name and background image), which can be easily manipulated, and platform-managed information (tweet timelines and timelines of friends' and followers' tweets), which is much harder to exploit. These algorithms also use more cues than any user could keep in

mind and combine them in rules without an intuitive, heuristic structure.

Social scientists have long touted the benefits of algorithms for overcoming human judgment and decision-making limitations—and have lamented the limited uptake of such aids (e.g., Dawes, 1979; Meehl, 1954). That experience has been repeated with ML aids created to support decision making in contexts as diverse as criminal justice (Kluttz & Mulligan, 2019), investing (Andriosopoulos et al., 2019), and healthcare (Sutton et al., 2020). Even systems that demonstrably outperform humans have not always been readily accepted (Burton et al., 2020; Diab et al., 2011; Dietvorst et al., 2015). In addition to addressing the bot detection challenge, our results contribute to general understanding of task and individual features that affect algorithm uptake. After introducing the task and our dependent measures of performance, we present those variables and the predicted relationships. Unless described as exploratory, all hypotheses were preregistered. Please see Supplementary Materials (SM) Tables SM1 and SM2 for details.

Task

In each of 60 trials, participants examined a public Twitter persona profile, indicated whether they believed that it was created by a "bot" or a human, and rated their confidence in that choice, using a slide bar on a scale anchored at 50% (completely uncertain) and 100% (certain). Finally, participants indicated whether they would retweet a message from that persona if they agreed with its content. Participants were randomly assigned to a Control group (with no aid); a Reminder group, with a repetitive warning about bots accompanying each persona; and an Aid group, with a bot indicator score accompanying each persona. The repetitive warning served as an additional means to determine if the presence of any type of indicator, even one that provided no additional information about a profile's probability of being a bot, might increase vigilance in the task, relative to the presence of the bot indicator score which did provide new information. After completing these tasks, participants answered demographic questions, a survey of social media experience, and additional exploratory measures. Participants in all conditions did not receive any explicit feedback on whether their judgment was accurate or not.

Stimulus Selection

We selected currently active Twitter personas. We estimated each persona's probability of being a social bot based on the agreement of three independently developed and trained ML algorithms: Bot Hunter (Beskow & Carley, 2018a), Botometer (Davis et al., 2016), and Bot Sight (Kats, 2022). We used Bot Hunter's Tier 1 model as the basis of our bot indicator score because it uses features accessible to typical users. We then selected personas with concurring Botometer and Bot Sight scores. We created a suite of stimuli whose bot indicator scores were roughly uniformly distributed from very low (1%) to very high (99%). Bot Hunter scores had correlations of .926 with Botometer scores and .88 with Bot Sight scores. More details are provided in the Supplementary Materials.

Dependent Measures

Sensitivity. Participants' *sensitivity* is the correlation between their categorical judgments (bot, human) and the bot indicator score. More details on this calculation are provided in the Supplementary Materials.

Criterion. Participants' *criterion* is their tendency to treat ambiguous personas as bots or humans, reflecting their relative value for correctly identifying members of each class. More details on this calculation are provided in the Supplementary Materials.

System Properties

We compared the performance of participants randomly assigned to one of three groups. *Control* group participants saw just the personas in the Twitter profiles. *Aid* group participants also saw a bot indicator score near the persona's name. *Reminder* group participants saw a warning about bots in the same place as the bot indicator score for the Aid group. We had the following preregistered hypotheses. H1 and H3 reflect replication of results in our previous study (Kenny et al., In press).

(H1–sensitivity) 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–sensitivity) Aid group participants will have greater sensitivity than the Control and Reminder groups.
(H3–criterion) Participants will have an aversion to misidentifying humans as bots, responding "human" more often than "bot," despite the equal number of each in the stimulus set.
(H4–criterion) Participants in the Aid and Reminder groups will have lower thresholds for responding "bot," compared to the Control group, as both interventions caution users against uncritically accepting bots as humans.

Task Engagement and Fatigue. We include two variables in our prediction models: (a) *Task Order (TO)* to see if fatigue reduces sensitivity (Parasuraman & Davies, 1977; Warm et al., 2008) and (b) *Task Engagement (TE)*, based on attention checks, to see if more engaged participants perform better, as seen in Kenny et al. (In press) and elsewhere (e.g., Dewitt et al., 2015; Downs et al., 2010; Matthews et al., 2010). The attention checks were three well-known public images with profiles consistent with fame (Mike Pence, Kim Kardashian, and Elizabeth Warren). We do not expect the experimental manipulations to affect these relationships.

(H5) Participants' sensitivity will decline with task order.
(H6) Participants who answer more attention checks correctly will demonstrate greater sensitivity.
(H7) We expect no correlation between either task engagement and fatigue and participants' decision criteria. We did not expect participants' response threshold to show any relationship with their level of engagement and relative levels of fatigue.

Methods

Sample

We collected data in January 2022. In all, 924 participants 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., 2017). Eyal et al. (2021) compared several online behavioral research platforms 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, and follower count). Each participant received 30 "bot" personas and 30 "human" personas, in a unique random order. Three additional interspersed trials presented personas of public figures (Elizabeth Warren, Mitch McConnell, and Kim Kardashian) as attention checks.

Participants were randomly assigned to view the personas with no aid (Control), a warning (Reminder), or the bot indicator score (Aid). Reminder group participants received the caution, "Look for Bot Cues," similarly placed in each stimulus. In the Aid condition, participants saw colored bot/human indicator icons with an exact numerical prediction for each persona. The colors reflected the likelihood of a bot or human, as depicted in Figure 2 and adapted from Bot Sight's beta social bot indicator add-in (Kats, 2022). If participants relied solely on the aid and used 50% as their cutoff point, they would be scored as perfectly accurate.

After the trials, participants completed a demographic survey and questions regarding social media experience, analytical abilities, and political views, for exploratory analyses of individual differences (Sarno & Black, in press). As input to future research, but not analyzed here, we elicited their willingness to pay for "an automated social bot detection service" and their concern over social bots, for themselves and for society.

Analysis Plan

We used a generalized linear mixed-effects probit regression to assess trial-level effects, predicting

participants' probability of calling each persona a bot (Table 1). This approach uses the unobserved heterogeneity in the intercept and the slope to capture participants' criterion and sensitivity to the bot indicator, respectively, 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 probabilities 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 et al., 2016). More details are provided in the Supplementary Materials.

Our preregistered analyses examined the contributions of the experimental condition (Control, Aid, Reminder) and stimuli attribute (bot indicator score) in predicting whether a participant responded "bot." Task order and 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 more substantial 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. The Bot Indicator (BI) variable is the algorithm-derived probability of the stimulus being a bot in each model, used here as a surrogate for ground truth. The intercept reflects the criterion for the average

Figure 2. Examples of human/bot indicator icons across the color range.

Table 1. General Linear Mixed Effects Probit Regression Model Predicting the Probability of Judging a Persona to Be a Bot.

Dependent Variable ('Bot' Response)			
	Model		
Predictors	Estimate	CI	*p*
Intercept (Criterion)	.094	−.166–.354	.478
Bot Indicator (BI)	−.501	−.943–−.058	**.027**
Task Order (TO)	.001	−.000–.002	.188
Task Engagement (TE)	−.189	−.242–−.135	**< .001**
Group [Reminder]	.033	−.057–.123	.469
Group [Aid]	−.633	−.725–−.541	**< .001**
BI × TO	.000	−.002–.003	.707
BI × TE	.290	.198–.381	**< .001**
BI * Group [Reminder]	.069	−.082–.221	.370
BI * Group [Aid]	1.646	1.489–1.803	**< .001**
N		928	
Observations		55680	
Marginal R^2 / Conditional R^2		.171/.288	
AUC		.756	

The bolded numbers are the significance level, interpreted as *p* values (in the column heading).

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 BI score reflect greater sensitivity. Figure 3 shows functions predicting the probability of responding "bot" based on BI, with the covariates in the Table 1 model.

Results

Sample Demographics

In all, 928 individuals completed the study. Six were excluded for not following Prolific's verification procedures. One was excluded for completing the task too quickly (less than 2 minutes); three were excluded for taking too long (more than 3 hours).

The analyzed sample included 373 males, 536 females, 16 nonbinary, and 3 who preferred not to say; their ages ranged from 18 to 78 years old (median = 33; mean = 36). Of them, 714 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 education, 368 a high school degree or equivalent, 396 a bachelor's degree, 128 a master's degree, and 31 a doctorate. 409 reported being fully employed, 123 employed part-time, 89 unemployed and looking, 51 unemployed not looking, 124 students, 35 retired, 76 self-employed, and 21 unable to work. 328 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 and 75K."

Criterion, Bot Indicator, Task Engagement, and Fatigue

Average Sensitivity. Participants in all three conditions demonstrated sensitivity to the bot indicator score, as evidenced by the positive trend of the curve (Figure 3) and the significant BI × Test Condition coefficient (Table 1). However, there were significant differences across conditions ($p < .001$). The correlation seen in the rising slope of the probability of a participant responding "bot" as the bot indicator probability increases reflects this increase in sensitivity. Participants in the Aid condition showed greater sensitivity than those in the Control or Reminder conditions—who were indistinguishable from one another. The corresponding percentages of correct identifications were Control = 56.3%, Reminder = 57.3%, and Aid = 71.9%.

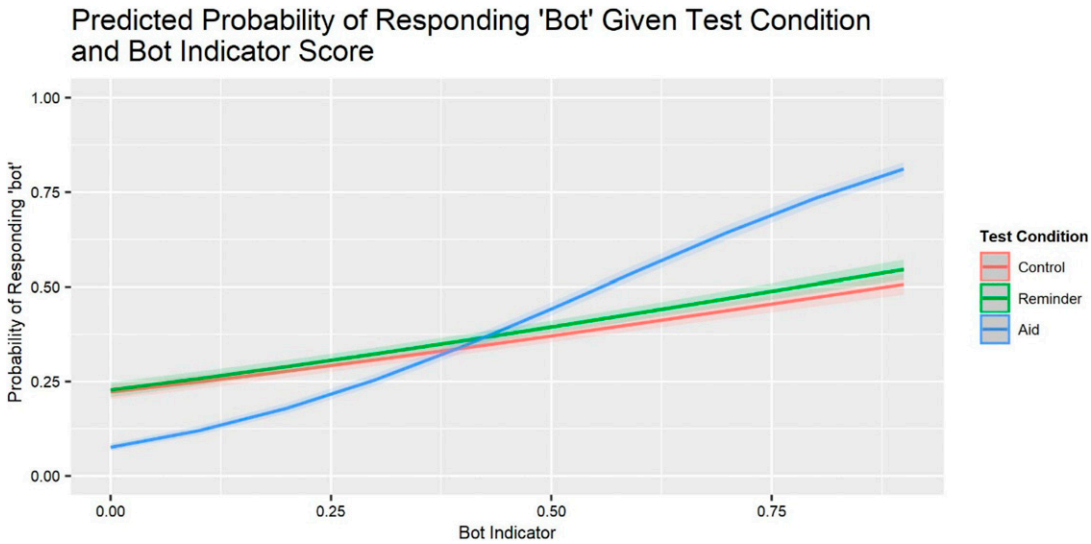

Figure 3. Predicted probability of responding "bot," given test condition and bot indicator score.

Average Criterion. Each group had a decision criterion reflecting a tendency to respond "human" (and hesitancy to respond "bot"), as seen in the intercepts greater than 0 (Figure 3) and negative bot indicator coefficient (Table 1). Participants in the Aid condition were more willing to call a persona a bot ($p < .001$) than participants in the other conditions, who were indistinguishable. This reflects a 6% increase in the probability of responding "bot" in the Aid condition, given the bot probability score, when holding all predictor variables constant.

Task Engagement and Fatigue. Task Order was unrelated to participants' criterion or sensitivity (BIxTO, Table 1). Higher Task Engagement was associated with greater willingness to respond "bot" ($p < .001$) and greater sensitivity ($p < .001$ (BIxTE, Table 1).

Confidence

Table 2 predicts participants' confidence given their test condition and response ("bot" or "human"), using a linear general mixed-effects model with the same predictors as above, but adding the classification judgment ("bot" or "human").

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. Confidence was similar across the test conditions (Control = 77.8%, Reminder = 77.1%, Aid = 78.5%). Overall, 61.9% of participants' judgments were accurate. Thus, participants were overconfident in their abilities, replicating a familiar finding in this novel setting (Canfield et al., 2016; Lichtenstein et al., 1982). However, overconfidence was much less in the Aid condition (6.6%) than the other two (Control = 21.5%; Reminder = 19.8%). Participants were less confident when they classified a persona as a bot, rather than a human ($p < .001$), a 6.7 percentage point difference.

Confidence was unrelated to task order (TO) or task engagement (TE). It decreased somewhat as the bot indicator score increased (BI × TE; Table 2). Figure 4 depicts the three-way (BI × SDT × Group) interaction. Aid condition participants' confidence was distinctly greater when their classification judgment was supported by a high or low bot indicator score.

Behavioral Responses

We used the same general linear mixed-effects modeling process as Table 1 to predict the probability that participants were willing to retweet content from a persona, if they agreed with its content. Participants were significantly less likely to retweet messages the higher a persona's bot indicator score, consistent with sensitivity (SM Table 1; Model 1). Willingness to retweet was higher in the Aid condition, suggesting that the bot indicator increased confidence in retweeting, and lower in the Reminder group, suggesting that a warning alone (without instructions for action) decreased confidence. The Aid group used the bot indicator score when deciding whether to retweet. Without the presence of the bot indicator, participants in the other conditions relied upon their judgments. Retweeting declined as the experiment progressed, suggesting fatigue. It was more common for participants with greater task engagement, implying that retweet decisions require effort in this experiment. Participants were more willing to tweet content from personas they judged to be humans, a form of manipulation check (Table SM1, Model 2). Participants were more willing to retweet, the greater their confidence in their response (bot or human) and the higher the bot indicator score (Table SM1, Model 3). Figure 5 depicts this pattern, pooling the three conditions, which did not differ.

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 (Glikson & Woolley, 2020; Hoff & Bashir, 2015; Lee & See, 2004; Pavlou, 2003). The present study examined how people use advice from a social bot detection algorithm, as measured by their sensitivity to social bots. Unlike many studies, we found that people made good use of the algorithm, increasing their detection sensitivity and shifting their decision criterion to a more cautious one, seemingly reflecting a better understanding of the threat posed by social bots. Providing a reminder with each test stimulus had no effect, indicating that the content of the aid, and not just its presence, made the

Table 2. Linear Mixed-Effects Model Predicting Self-Rated Bot Detection Confidence Scores.

Dependent Variable Self Rated Confidence (50% - 100%)			
	Model		
Predictors	Estimate	CI	*p*
(Intercept)	82.019	77.797–86.241	**< .001**
Bot Indicator (BI)	−2.277	−5.318–.763	.142
Task Order (TO)	.001	−.010–.013	.792
Task Engagement (TE)	−.267	−1.141–.608	.550
SDT Response ('bot' = 1)	−6.683	−7.451–−5.916	**< .001**
Group [Reminder]	−1.423	−2.917–.071	.062
Group [Aid]	3.375	1.889–4.862	**< .001**
BI × TO	.005	−.014–.024	.593
BI × TE	−.617	−1.227–−.006	**.048**
BI × SDT	9.429	8.156–10.703	**< .001**
BI × Group [Reminder]	.563	−.688–1.815	.378
BI × Group [Aid]	−9.959	−11.289–−8.630	**< .001**
SDT × Group [Reminder]	1.417	.332–2.501	**.010**
SDT × Group [Aid]	−9.380	−10.518–−8.242	**< .001**
BI × SDT × Group [Reminder]	−.581	−2.386–1.223	.528
BI × SDT × Group [Aid]	21.843	19.949–23.737	**< .001**
N		928	
Observations		55680	
Marginal R^2/Conditional R^2		.032/.372	

The bolded numbers are the significance level, interpreted as *p* values (in the column heading).

Figure 4. Relationship between the bot indicator score of each stimulus and the self-rated confidence for each participant response and experimental condition.

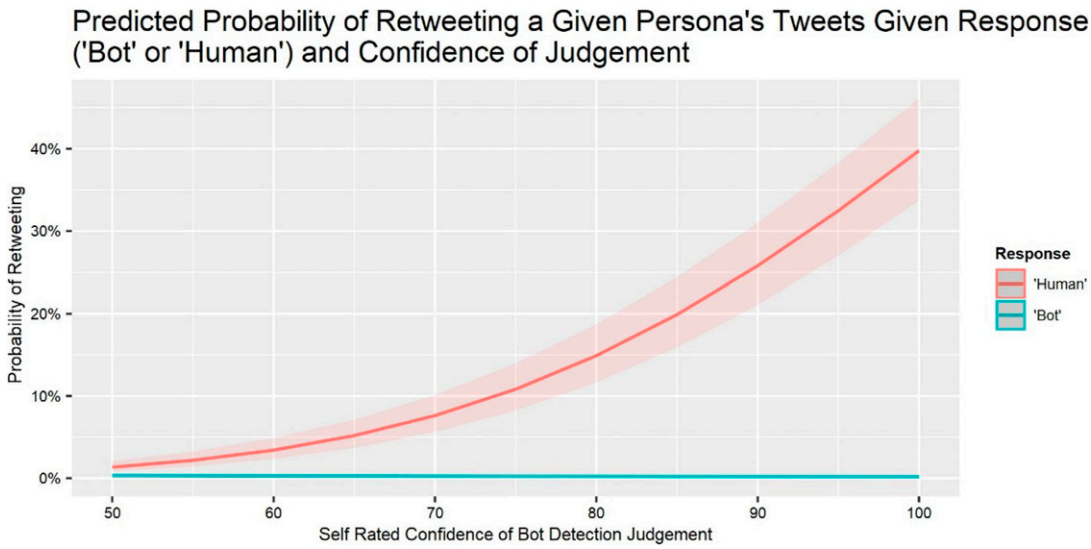

Figure 5. Relationship between self-rated confidence in bot detection judgment and the probability of retweeting a given persona's content for each participant response and experimental condition.

difference. The failure of warnings is a common result (Fischhoff, 1982; Milkman et al., 2009; Sharevski et al., 2022). Confidence was similar in the warning and control groups, indicating that the warning did not give a false sense of confidence.

Participants in the Control and Reminder conditions had modest sensitivity to the bot indicator score, supporting H1. The aid, showing participants the bot indicator score, significantly increased their sensitivity, supporting H2. Participants in the Control condition had decision criteria biased toward assuming that personas were humans, replicating our previous research (Kenny et al., In press), supporting H3, and consistent with a general tendency to treat claims as true (Sarno & Black, in press). Participants in the Aid condition were more cautious about treating personas as humans; participants in the Reminder condition were similar to Controls, partially supporting H4. Exploratory analyses found that the aid improved performance without increasing confidence, thereby reducing overconfidence. Thus, the aid improved performance (sensitivity), increased caution (criterion), and reduced overconfidence, while the reminder had no effect. Participants rarely were willing to retweet content, unless they agreed with it and were confident that it came from a human (Figure 5), indicating the importance of these abilities.

Experiment 2

Experiment 1 found that individuals could use bot indicator scores, based on ML algorithms, to improve their judgments of Twitter personas and adopt more cautious decision criteria. Experiment 2 examines the robustness of this result by replicating the Aid and Control conditions of Experiment 1, but not the failed Reminder condition. It also evaluates a training intervention based on the algorithms, suited to settings where they are not available.

Lack of availability can happen for various reasons. Although algorithms can help detect diseases, their adoption has been slow due to preferences for human advice (Dietvorst et al., 2015; Buck et al., 2022; Gaube et al., 2021). Even algorithms that are accepted as valid may be avoided if users fear that they will lead to complacency, dependence, and deskilling (Bahner et al., 2008; Dwivedi et al., 2021). Social media platforms may discourage the use of bot detection services that could reveal their algorithms, add costs that cannot be passed on to users, or cause embarrassment (e.g., by mislabeling personas) (Constantin, 2022; Jagielski et al., 2018).

Experiment 2 develops and evaluates an intervention designed to address these concerns: a video that trains users in heuristic rules for identifying bots. Those heuristics do not depend on

social media platforms to provide (or allow) automated bot detectors. They are always available, even for new personas with too little data for algorithms to analyze. They require ongoing user engagement, reducing the risk of complacency and deskilling.

As with any intervention, success depends on the quality of execution. Heuristics must be valid predictors of whether personas are social bots. The training must convey these heuristics in intuitively meaningful terms. Users must have the confidence, skills, and cues needed to apply them. The next section describes our development process.

Training for Social Bot Detection

Professionals refine their heuristics through experience, informed by quality feedback (Kahneman & Klein, 2009). However, as Simon and Chase (1973) proposed, an individual's experience becomes useful only when it leads to the pattern recognition essential to heuristic search and inference (Ericsson, 2017; Hanushek & Rivkin, 2010). In that light, we patterned our intervention after one that has helped physicians refine their heuristics for trauma triage (Mohan et al., 2017, 2018). Its foundation rests on heuristic rules that (a) have diagnostic value and (b) can be easily integrated with natural ways of thinking. As a result, we sought cues in publicly visible Twitter profiles with sufficient bot signal strength to have the potential for training and a memorable, heuristic organizing principle.

We conjectured that one such principle is that social bot creators want to reach as many people as possible as fast as possible. We tested this conjecture with analytical tests evaluating the predictive value of related cues readily found on a Twitter persona profile page. Those cues included having (a) many Tweets, relative to the age of the account, trying to amplify narratives quickly and (b) many followed personas, trying to enlist followers. The Supplementary Materials describe these tests of cues' ability to predict Bot Hunter's Tier 1 results. The first test trained four machine learning algorithms using four Tier 1 platform-managed features that suggest a rapid amplification strategy: account age, total number of tweets, number of followed accounts, and number of following accounts. We found that the best algorithm using these four features had an accuracy of 85.8%. The second test applied the same four machine learning algorithms to the visible Tier 1 features not used in the first test. It found that they had much less predictive value, with the best model having predictive accuracy of only 68.5%.

These tests used a publicly available training set. We tested the applicability of the resulting four-cue algorithm (from the first test) to our study by assessing its accuracy for the 60 personas used in our experiments. It was 90% accurate, giving us confidence that its four cues had sufficient signal strength to inform participants' judgments—if people could use them. Our training video (described below) had this goal.

Predicted Relationships

Participants were randomly assigned to one of three groups. All performed the same 60-persona task of Experiment 1, after watching a new 40-s introductory video about the threat posed by social bots. The Aid group was the same as that in Experiment 1. Before completing the tasks, the Training group watched a two-part instructional video on social bot detection. The Control group went directly to the persona evaluation task, as before.

We had the same two primary dependent measures: sensitivity and criterion. We made the following predictions, which were preregistered, unless otherwise noted. We had no predictions regarding the size of effects in the two intervention groups (Aid, Training). H1, H2, H4, and H5 predict replication of results in Experiment 1. H1 and H4 also reflect replications of results in Kenny et al. (In press).

Sensitivity

(H1) The probability of participants responding "bot" will increase as the bot signal increases, for all three conditions.

(H2) Participants in the Aid condition will perform better than those in the Control condition.

(H3) Participants in the Training condition will have greater sensitivity than those in the control group.

Criterion

(H4) Control group participants will have greater aversion to mistaking a human for a bot than vice versa.
(H5) Participants in the Aid group will be less averse to mistaking a human for a bot than those in the Control group.
(H6) As the video emphasizes social bot developers' motives, Training group participants will be more averse to mistaking a bot for a human than are Control group participants.

Task Engagement and Fatigue. We included the same two control variables as in Experiment 1, with the same predictions, despite mixed results there and in Kenny et al. (In press).

(H7) Participants' sensitivity will decrease with Task Order, reflecting fatigue.
(H8) Participants with higher Task Engagement scores will have greater sensitivity and show greater improvements in sensitivity with training.
(H9) Neither control variable will be related to participants' decision criteria.

Methods

Sample

We collected data in March 2022. We recruited 924 participants 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

Experiment 2 had the same design as Experiment 1, except for the change in test conditions and the new 40-s introductory video, defining social bots and highlighting their threat. Our analysis plan was the same as well.

Training Protocol

Participants in the Training condition watched two additional videos, following the introductory video that all participants saw. The first, lasting approximately 90 seconds, explained the importance of identifying bots and ignoring the content of their messages. It drew attention to two cues: the persona's total number of tweets and the account's age. It said, "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."

It then drew attention to two cues related to the size of the persona's social network: how many people the persona is following and how many people were following the persona. It said, "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."

Finally, Training group participants watched a two-minute video that led them through four examples illustrating use of these cues. It concluded with two comprehension questions intended to keep then engaged. Almost all answered both questions correctly.

Results

Sample Demographics

In all, 982 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 6 minutes), and two for taking too long (more than 3 hours).

The analyzed sample included 976 participants, 433 males, 531 females, 10 nonbinary, and 2 preferred not to say; age ranged from 18 to 84 years old (median = 36; mean = 38). Among them, 783 reported being White, 55 Hispanic or Latino, 50 Black or African American, 5 Native American, 64 Asian or Pacific Islander, and 19 Other. For education, 9 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. For employment status, 465 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. Overall, 388 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 and 50K.

Dependent Measures

Sensitivity. Table 3 and Figure 6 show patterns like those in Experiment 1 for the Control and Aid groups. All three groups were sensitive to the bot indicator score (BI). The Aid group was significantly more sensitive than the Control group (BIxAid), as was the Training group (BIxTraining), although somewhat less than the Aid group. The correlation reflected in the rising slope of the probability of a participant responding "bot" as the bot indicator probability increases represents this increase in sensitivity.

Criterion. As in Experiment 1 and Kenny et al. (In press), Control group participants exhibited a bias toward labeling stimuli as humans. As in Experiment 1, that tendency was significantly less with the Aid group (Group [Aid]). The Training group showed less of that bias as well (Group [Training]).

Task Engagement and Fatigue. Sensitivity increased with task order (BIxTO), suggesting improvement with practice, rather than the predicted fatigue. Sensitivity increased with Task Engagement (BIxTE). The bias toward responding "human" decreased with both TO and TE (we had no prediction).

Confidence

As in Experiment 1, participants expressed moderate confidence in their SDT (bot/human) judgments. Pooling groups and stimuli, M = 78.7% (*SD* = 15.1%), on the 50%–100% scale. Given their 65.1% accuracy rate, participants were overconfident overall (78.7%–65.1% = 13.6%). As seen in Table 4, confidence was significantly higher in the Aid group (M = 77.9%) and the Training group (M = 81.9%) than in the Control group (M = 76.2%). However, as accuracy was much higher in the Aid group (70.6%) and Training group (67.9%) than in the Control group

Table 3. General Linear Mixed Effects Probit Regression Model Predicting the Probability of Judging a Persona to Be a Bot.

Dependent Variable ('Bot' Response)			
	Model		
Predictors	Estimate	CI	*p*
Intercept (Criterion)	.738	.451–1.025	**< .001**
Bot Indicator (BI)	−1.205	−1.725–−.685	**< .001**
Task Order (TO)	−.002	−.003–−.001	**.001**
Task Engagement (TE)	−.276	−.336–−.216	**< .001**
Group [Training]	−.519	−.616–−.422	**< .001**
Group [Aid]	−.562	−.659–−.465	**< .001**
BI × TO	.003	.001–.005	**.004**
BI × TE	.419	.311–.527	**< .001**
BI * Group [Training]	.990	.815–1.165	**< .001**
BI * Group [Aid]	1.520	1.343–1.696	**< .001**
N		976	
Observations		58560	
Marginal R^2/Conditional R^2		.205/.328	
AUC		.771	

The bolded numbers are the significance level, interpreted as *p* values (in the column heading).

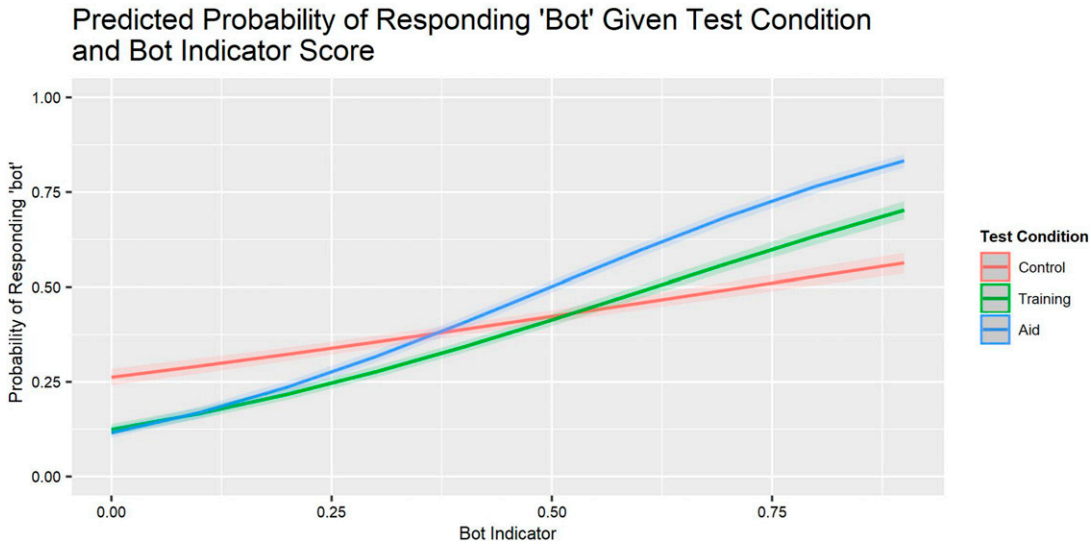

Figure 6. Relationship between the bot indicator score of each stimulus and the probability of responding 'bot' across experimental conditions.

Table 4. Linear Mixed-Effects Model Predicting Self-Rated Bot Detection Confidence Scores.

Dependent Variable Self-Rated Confidence (50% - 100%)			
	Model		
Predictors	Estimate	CI	*p*
(Intercept)	80.721	76.484 to 84.959	**< .001**
Bot Indicator (BI)	−3.057	−6.566 to .452	.088
Task Order (TO)	.003	−.007 to .014	.534
Task Engagement (TE)	−.568	−1.452 to .317	.208
SDT Response ("bot" = 1)	−4.113	−4.835 to −3.391	**< .001**
Group [Training]	7.493	6.043 to 8.944	**< .001**
Group [Aid]	4.626	3.171 to 6.081	**< .001**
BI × TO	−.008	−.027 to .010	.384
BI × TE	−.211	−.926 to .504	.563
BI × SDT	8.076	6.854 to 9.298	**< .001**
BI × Group [Training]	−2.146	−3.515 to −.777	**.002**
BI × Group [Aid]	−10.449	−11.913 to −8.985	**< .001**
SDT × Group [Training]	−3.281	−4.388 to −2.175	**< .001**
SDT × Group [Aid]	−9.285	−10.339 to −8.230	**< .001**
BI × SDT × Group [Training]	2.194	.363 to 4.024	**.019**
BI × SDT × Group [Aid]	21.544	19.731 to 23.356	**< .001**
N		928	
Observations		55680	
Marginal R^2/Conditional R^2		.032/0.372	

The bolded numbers are the significance level, interpreted as *p* values (in the column heading).

(56.4%), their overconfidence was less: Aid = 7.3%, Training = 14.0%, and Control = 19.8%.

Other patterns also resembled those in Experiment 1. Confidence was unrelated to Task Order or Task Engagement. It was lower when participants responded "bot," rather than "human" (4.1% mean difference, $p < .001$). There was a three-way interaction (BI × SDT × Group) similar to that in Table 2 and Figure 4. As seen in Figure 7, confidence was much more sensitive to the bot indicator score in the Aid group than in the Control group, with the Training group in between.

Behavioral Responses

As in Experiment 1, participants would rarely retweet messages, even when they agreed with the content: Aid = 14%, Control = 15%, and Training = 22%. Consistent with Experiment 1 and Figure 5, willingness to retweet increased greatly with confidence for personas seen as humans. Table SM2 shows results of exploratory analyses using general linear mixed-effects regressions to predict retweet probability.

Discussion

The Control group replicated results from Experiment 1 and Kenny et al. (In press). Under these experimental conditions, meant to simulate the rapid assessments made in online environments, these unaided individuals were somewhat sensitive to whether personas are social bots, overconfident in their discrimination ability, and biased toward assuming that personas are humans rather than bots. Providing a bot indicator score for each persona significantly increased sensitivity, reduced overconfidence, and shifted the decision criterion, toward being more cautious about potentially treating a bot as a human. We found that people will and can use this algorithmic aid.

The training video condition was designed for situations where algorithmic interventions were unavailable. It described and demonstrated two cues with predictive value, as established by our tests, and potentially with the intuitive resonance needed to become part of Twitter users' natural heuristics. The training video improved sensitivity and reduced overconfidence, although seemingly somewhat less than did the aid. It did not affect participants' decision criteria toward greater caution, despite warning them about the motives of social bot creators.

Pooling across groups, participants' behavioral responses were generally consistent with their judgments of the personas. They were reluctant to retweet messages, even if they agreed with the content, unless they were confident that it was from a human. Their confidence and agreement

Figure 7. Relationship between the bot indicator score of each stimulus and the self-rated confidence for each participant response and experimental condition.

increased when the bot indicator score was very high or very low, even if they could not see it.

Conclusions and Applications

We found evidence for two strategies that can improve social bot detection: a bot indicator score, provided with each tweet, and a brief training video. Both interventions increased users' sensitivity to whether personas were humans or bots. The bot indicator score aid shifted users' criterion toward being more cautious about treating a bot as a human. The video training did not affect the criterion, despite emphasizing the manipulative strategies of social bot developers. In all conditions, participants were reluctant to retweet a message, even if they agreed with its content, unless they were confident that it came from a human. Thus, any intervention that improves sensitivity to social bots should reduce their impact.

In the current set of experiments, participants did not receive feedback about their performance as they completed the task. When presented with probabilities of a persona being a social bot or not, participants made use of this information as evidenced by increased sensitivity and decreased willingness to share content. When trained prior to beginning the task, participants demonstrated that they retained the trained detection heuristics. The improved sensitivity, and increased success in detecting social bots, therefore, may not be contingent upon receiving explicit feedback.

The training video, which adapts an approach from Mohan et al. (2017, 2018) is, we believe, the first of its kind in this domain. Its success suggests pursuing its strategy further. First, identify statistically valid cues, as we did with our tests of the predictive value of algorithms using the cues. Then, help users incorporate those cues into their intuitive heuristics by explaining their rationale, giving users a mental model of the system (Bruine de Bruin & Bostrom, 2013). Here, that meant explaining that social bot developers seek to amplify their narrative as quickly as possible. As a result, excessive tweeting over a short period of time suggests a bot, as does a high follow-to-follower ratio, because social bot developers seek to create an extensive social network. Thus, the training strategy requires understanding both the system and the users.

Heuristic training can also empower users when algorithms are unavailable, which may be especially important when bots are too new to be characterized statistically, as part of adversarial social bot developers' attempts to avoid detection (Cresci, 2020; Orabi et al., 2020). In this game of cat and mouse, between social bot developers and their targets, heuristics may prove more robust than algorithms, while keeping users engaged. Both training and algorithms 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. Such collaborations are a hallmark of human factors.

Limitations and Future Work

Our conclusions depend on the accuracy of the three machine learning algorithms that we used to evaluate performance, in the absence of ground truth indicators of whether personas were bots or human. Like other machine learning models, Botometer, Bot Hunter, and Bot Sight may have been trained on unrepresentative and mislabeled training sets, with unknown effects on our results. Our strategy for addressing these risks is to use stimuli with similar scores on the three systems, hoping that their respective training sets and computational procedures would not share biases. This is a fundamental challenge for tasks where automation is employed, but a ground truth is unavailable. As social bot developers advance the sophistication of their automated tools, even the most rigorously developed algorithms will lose validity over time, as the information environment changes.

Our conclusions also depend on the external validity of our experimental task. As with other simulated experiences (Aiello et al., 2012; Wald et al., 2013), its validity depends on how well it evokes real-world behavior. We used actual personas and set a pace akin to the rapid evaluation typical of Twitter use. However, we did not provide access to the persona profile pages that users might examine, when suspicious of a persona. As a result, we might have underestimated users' abilities.

We also had a higher proportion of social bots (50%) than in everyday Twitter life. As a result, we may have increased participants' tendency to identify them as human (Varol et al., 2017). Within that constraint, the stimuli had a roughly rectangular distribution of bot indicator scores. Thus, they varied in difficulty for participants and in the diagnosticity of the scores, both as feedback for participants and for evaluating their performance. Those evaluations could be done more confidently, assuming the validity of the scores, by excluding some fraction with scores close to chance. That exclusion would change the parameter estimates in our models, but would not, we believe, materially change the patterns or relative effectiveness of the proposed interventions.

Our task may have enhanced performance, by creating a test environment, or reduced it, by having only intrinsic incentives. Performance generally increased with task engagement. We cannot know, without retesting, how long the effects of the video training would last or whether continuing exposure to the bot indicator 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 et al., 2022).

The practical implications depend upon the corollary of our task and how average social media users might expect to act when faced with uncertainty about a persona's origin. When attention is limited (Hodas & Lerman, 2012; Lerman et al., 2013), social bots may go unnoticed by average users. However, if users are suspicious and investigate persona details on their profile page, or if users are alerted to a persona's probability of being a bot, as was the case in our experiments, then these studies suggest both increased detection of bots and a corresponding reduction in users' willingness to share their content. In other words, the aid may serve as an alert or it may serve as a type of feedback. More research is needed to understand the cognitive mechanism.

Conclusion

Social bot developers are becoming increasingly sophisticated at mimicking human personas and manipulating users' commercial and political behavior. Our findings have identified two possible interventions that can improve online users' detection of social bots and reduce the likelihood of the sharing of social bot content: algorithmic bot indicator scores, accompanying each tweet, and video training, creating heuristic mental models of how to detect and manage the risks.

Key Points

- We demonstrate a platform for evaluating Twitter users' performance in distinguishing human and social bot personas.
- We identify two interventions that increased sensitivity in our simulated environment: an algorithmic aid, providing bot indicator scores for each tweet, and a training video, focused on heuristic mental models for bot detection.
- The aid increased users' aversion to mistaking a human for a bot; the training video did not.

Acknowledgments

Funding for this research was provided by the U.S. Army, Advanced Strategic Planning and Policy Program, Goodpaster Fellowship. We thank Dave Beskow for access to Bot Hunter and his early advice on this endeavor, Jonathan Kats for his feedback regarding Bot Sight, and Megan Kenny, as an independent coder. The views expressed are those of the authors. The present research was conducted as part of Ryan Kenny's doctoral dissertation, in the Department of Engineering and Public Policy, Carnegie Mellon University, submitted May 2022 and titled, "Why some are better than others at detecting social bots: Comparing baseline performance with aids and training."

ORCID iDs

Baruch Fischhoff https://orcid.org/0000-0002-3030-6874

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

Supplemental Material

Supplemental material for this article is available online.

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.

Proceedings of the International AAAI Conference on Web and Social Media, *6*(1), 10–17. https://doi.org/10.1609/icwsm.v6i1.14236

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. https://doi.org/10.1080/01605682.2019.1595193

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

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. https://doi.org/10.1016/j.ijhcs.2008.06.001

Batailler, C., Brannon, S. M., Teas, P. E., & Gawronski, B. (2022). A signal detection approach to understanding the identification of fake news. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *17*(1), 78–98. https://doi.org/10.1177/1745691620986135

Beskow, D. & Carley, K. (2018a). Bot-hunter: A tiered approach to detecting and 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, USA, 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.

Bond, C. F. Jr. & DePaulo, B. M. (2006). Accuracy of deception judgments. *Personality and Social Psychology Review: An Official Journal of the Society for Personality and Social Psychology, Inc*, *10*(3), 214–234. https://doi.org/10.1207/s15327957pspr1003_2

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

Bruine de Bruin, W. & Bostrom, A. (2013). Assessing what to address in science communication. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(Suppl 3), 14062–14068. https://doi.org/10.1073/pnas.1212729110

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), Article e28916. https://doi.org/10.2196/28916

Burton, J. W., Stein, M. K., & Jensen, T. B. (2020). A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making*, *33*(2), 220–239. https://doi.org/10.1002/bdm.2155

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

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). Austin, Texas, USA, 6–10 December 2010.

Cohen, B. (2023, February 16). *Elon Musk's new enemy: An Army of good bots*. Wall Street Journal.

Constantin, L. (2022). *How data poisoning attacks corrupt machine learning models* (Vol. 14). CSO. Retrieved from https://www.csoonline.com/article/3613932/how-data-poisoningattacks-corrupt-machine-learning-models.html

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

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

Davis, C. A., Varol, O., Ferrara, E., Flammini, A., & Menczer, F. (2016). BotOrNot: A system to evaluate social bots. In WWW'16 Companion: Proceedings of the 25th International Conference Companion on World Wide Web (pp. 273–274). Montréal, QC, Canada, April 2016. 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), 571–582. https://doi.org/10.1037/0003-066x.34.7.571

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

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

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. https://doi.org/10.1111/j.1468-2389.2011.00548.x

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

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). Atlanta, GA, USA, 10–15 April 2010.

Dwivedi, Y. K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., Duan, Y., Dwivedi, R., Edwards, J., Eirug, A., Galanos, V., Ilavarasan, P. V., Janssen, M., Jones, P., Kar, A. K., Kizgin, H., Kronemann, B., Lal, B., Lucini, B., & 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. https://doi.org/10.1016/j.ijinfomgt.2019.08.002

Ericsson, K. A. (2017). Expertise and individual differences: The search for the structure and acquisition of experts' superior performance. *WIREs Cognitive Science*, *8*(1–2), Article e1382. https://doi.org/10.1002/wcs.1382

Eyal, P., David, R., Andrew, G., Zak, E., & Ekaterina, D. (2021). Data quality of platforms and panels for online behavioral research. *Behavior Research Methods*, *54*(4), 1–20. https://doi.org/10.3758/s13428-021-01694-3

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

Fischhoff, B. (1982). Debiasing. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 422–444). Cambridge University Press.

Gaube, S., Suresh, H., Raue, M., Merritt, A., Berkowitz, S. J., Lermer, E., Coughlin, J. F., Guttag, J. V., Colak, E., & Ghassemi, M. (2021). Do as AI say: Susceptibility in deployment of clinical decision-aids. *NPJ Digital Medicine*, *4*(1), 31. https://doi.org/10.1038/s41746-021-00385-9

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.

Glikson, E. & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *The Academy of Management Annals*, *14*(2), 627–660. https://doi.org/10.5465/annals.2018.0057

Green, D. M. & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Hanushek, E. A. & Rivkin, S. G. (2010). Generalizations about using value-added measures of teacher quality. *The American Economic Review*, *100*(2), 267–271. https://doi.org/10.1257/aer.100.2.267

Harari, Y. (2023, March 27). *You can have the blue pill or the red pill and we're out of blue pills* (p. 18). New York Times.

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. https://doi.org/10.1177/0093650214534974

Hodas, N. O. & Lerman, K. (2012, September). How visibility and divided attention constrain social contagion. In 2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Conference on Social Computing (pp. 249–257). IEEE.

Hoff, K. A. & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, *57*(3), 407–434. https://doi.org/10.1177/0018720814547570

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

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

Jagielski, M., Oprea, A., Biggio, B., Liu, C., Nita-Rotaru, C., & Li, B. (2018, May). Manipulating machine learning: Poisoning attacks and countermeasures for regression learning. In 2018 IEEE Symposium on Security and Privacy (SP) (pp. 19–35). San Francisco, CA, USA, 20–24 May 2018.

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

Kahneman, D. & Klein, G. (2009). Conditions for intuitive expertise: A failure to disagree. *American Psychologist*, *64*(6), 515–526. https://doi.org/10.1037/a0016755

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

Kenny, R., Fischhoff, B., Davis, A. L., Carley, K. M., & Canfield, C. (In press). Duped by bots: Why some are better than others at detecting fake social media. *Human Factors*, 187208211072642. https://doi.org/10.1177/00187208211072642

Kluttz, D. N. & Mulligan, D. K. (2019). Automated decision support technologies and the legal profession. *Berkeley Technology Law Journal*, *34*(3), 853–890.

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). Raleigh, NC, USA, 26–30 April 2010. 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. https://doi.org/10.1518/hfes.46.1.50_30392

Lerman, K., Jain, P., Ghosh, R., Kang, J. H., & Kumaraguru, P. (2013, May). Limited attention and centrality in social networks. In 2013 IEEE International Conference on Social Intelligence and Technology (pp. 80–89). State College, PA, USA, 08–10 May 2013.

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 306–334). Cambridge University Press. https://doi.org/10.1017/CBO9780511809477.023

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

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

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

Meaker, M. (2022). This student's side project will help to decide Musk vs. Twitter. https://www.wired.com/story/musk-twitter-botometer/ (accessed 6/12/23).

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

Milkman, K. L., Chugh, D., & Bazerman, M. H. (2009). How can decision making be improved? *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *4*(4), 379–383. https://doi.org/10.1111/j.1745-6924.2009.01142.x

Mohan, D., Farris, C., Fischhoff, B., Rosengart, M. R., Angus, D., Yealy, D., Wallace, D., & Barnato, A. (2017). Testing the efficacy of a video game vs. a traditional education program at improving physician decision making in trauma triage: A randomized controlled trial. *BMJ*, *359*, j5416. https://doi.org/10.1136/bmj.j5416

Mohan, D., Fischhoff, B., Angus, D. C., Rosengart, M. R., Wallace, D. J., Yealy, D. M., Farris, C., Chang, C.-C. H., Kerti, S., & Barnato, A. E. (2018). Serious games may improve physician heuristics in trauma triage. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(37), 9204–9209. https://doi.org/10.1073/pnas.1805450115

Orabi, M., Mouheb, D., Al Aghbari, Z., & Kamel, I. (2020). Detection of bots in social media: A systematic review. *Information Processing and Management*, *57*(4), 102250. https://doi.org/10.1016/j.ipm.2020.102250

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

Palan, S. & Schitter, C. (2018). Prolific. ac—a subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, *17*, 22–27. https://doi.org/10.1016/j.jbef.2017.12.004

Parasuraman, R. & Davies, D. R. (1977). A taxonomic analysis of vigilance performance. *Vigilance* (pp. 559–574). Springer.

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. https://doi.org/10.1080/10864415.2003.11044275

Peer, 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. https://doi.org/10.1016/j.jesp.2017.01.006

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

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

Sarno, D. M. & Black, J. Who gets caught in the web of lies? Understanding susceptibility to phishing emails, fake news headlines, and scam text messages. *Human Factors*, 187208231173263. (in press). https://doi.org/10.1177/0018720823117326

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*, *28*(1), 85–99. https://doi.org/10.1037/xap0000410

Satariano, A. & Mozur. (2023, February 9). *The people onscreen are fake. The disinformation is real* (p. 3). New York Times.

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

Sharevski, F., Alsaadi, R., Jachim, P., & Pieroni, E. (2022). Misinformation warnings: Twitters soft moderation effects on COVID-19 vaccine belief echoes. *Computers & Security*, *114*, 102577. https://doi.org/10.1016/j.cose.2021.102577

Simon, H. A. & Chase, W. G. (1973). Skill in chess. *American Scientist*, *61*(4), 394–403. https://www.jstor.org/stable/27843878

Stanislaw, H. & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers: A Journal of the Psychonomic Society, Inc*, *31*(1), 137–149. https://doi.org/10.3758/bf03207704

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.

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), 17. https://doi.org/10.1038/s41746-020-0221-y

Thompson, S. A. (2023, May 20). *A. I.-generated content discovered on news sites, content farms and product reviews* (p. 5). New York Times.

Timberg, C. & Dwoskin, E. (2018). Twitter is sweeping out fake accounts like never before, putting user growth at risk. 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/

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. https://doi.org/10.1007/s42001-020-00087-4

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*, *11*(1), 280, https://doi.org/10.1609/icwsm.v11i1.14871

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). San Francisco, CA, USA, 14–16 August 2013. IEEE.

Warm, J. S., Parasuraman, R., & Matthews, G. (2008). Vigilance requires hard mental work and is stressful. *Human Factors*, *50*(3), 433–441. https://doi.org/10.1518/001872008X312152

Warren-West, L. S. & Jackson, R. C. (2020). Seeing the bigger picture: Susceptibility to, and detection of, deception. *Journal of Sport & Exercise Psychology*, *42*(6), 463–471. https://doi.org/10.1123/jsep.2020-0040

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. https://doi.org/10.1145/3373464.3373475

Yang, K. C., Ferrara, E., & Menczer, F. (2022). Botometer 101: Social bot practicum for computational social scientists. *Journal of Computational Social Science*, *5*(2), 1511–1528. https://doi.org/10.1007/s42001-022-00177-5

Author Biographies

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 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 is a 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 is an 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 (BS 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.

Casey Canfield is an assistant professor in Engineering Management & Systems Engineering at Missouri University of Science & Technology. She has a BS in engineering: systems from Olin College of Engineering and a PhD 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.