Asymmetrical Perceptions of Partisan Political Bots

Authors: Harry Yaojun Yan[1,2], Kai-Cheng Yang[2], Filippo Menczer[2], James Shanahan[1]

Affiliations: 1. The Media School, Indiana University-Bloomington

2. Luddy School of Informatics, Computing, and Engineering

This manuscript has been accepted for publication at *New Media & Society*

Abstract

Political bots are social media algorithms that impersonate political actors and interact with other users, aiming to influence public opinion. This study examines perceptions of bots with partisan personas by conducting an online experiment ($N = 656$) that examines the ability to differentiate bots from humans on Twitter. We explore how characteristics of the experiment participants and the profiles being evaluated bias recognition accuracy. Our analysis reveals asymmetrical partisan-motivated reasoning, in that conservative profiles appear to be more confusing and Republican participants perform less well in the recognition task. Moreover, Republican users are more likely to confuse conservative bots with humans, whereas Democratic users are more likely to confuse conservative human users with bots. We discuss implications for how partisan identities affect motivated reasoning and how political bots exacerbate political polarization.

Keywords: Political bots, Social media, Online human-bot interaction, Partisan-motivated reasoning, Political polarization

Asymmetrical Perceptions of Partisan Political Bots

In January 2018, a *New York Times*' investigation exposed the business of manufacturing social media popularity. One Florida company called *Devumi* alone produced and sold social bots accounting for over 200 million fake Twitter followers. Social bots are defined as computer algorithms that use software to produce content automatically or semi-automatically, trying to emulate and possibly alter human behavior (Ferrara, Varol, Davis, Menczer, & Flammini, 2016). Although not all bots are created with malicious intent, bots designed to sway public opinion and potentially change political behaviors have been found to be damaging and prevalent, particularly since the 2016 U.S. presidential campaign (Badawy, Ferrara, & Lerman, 2018; Bessi & Ferrara, 2016; Shao et al., 2018a).

The emergence of social bots, including political bots, has facilitated the distribution of false information in at least two ways. First, evolving bot algorithms that mimic online human behavior make even experienced users vulnerable (Wagner, Mitter, Körner, & Strohmaier, 2012). Bots can monitor user traffic flow and follow circadian rhythms to maximize the visibility of their content (Ferrara et al., 2016). Second, network effects can inflate the popularity of certain issues (Ferrara et al., 2016). So-called "botnets"—coordinated groups of inauthentic accounts—can do everything regular users do to generate influence, but faster, at a more massive scale, and at lower cost (Boshmaf, Muslukhov, Beznosov, & Ripeanu, 2013). When a user observes a tweet or an account profile along with their social influence indicators (number of retweets, replies, likes, friends, and followers), it can be virtually impossible to efficiently discern if the indicators are genuine or inflated by bots, regardless of the legitimacy of the tweet or account.

In comparison to generic social bots, political bots that have explicit partisan personas can be even more deceptive in the contemporary political climate. First, the appearance of political bots can be more strategically organized, and the attacks more targeted (Stella, Ferrara, & De Domenico, 2018). The deployment of political bots has been observed during election cycles (Bessi & Ferrara, 2016), swamping voters already overwhelmed by a dramatic increase in political information. With a polarized ideological gap that is at its historical peak in the US (Kiley, 2017), political bots are likely to exploit the biased perceptions of people with different party identifications, making them more vulnerable to manipulation.

In light of a Pew Research survey showing an increased public awareness of social bots (Stoking & Sumida, 2018), the main goal of this study is to explore and understand the human capability to distinguish political bots from human users. This task presents both the challenge of recognition (Gardiner, Ramponi, & Richardson-Klavehn, 2002) and that of decision under uncertainty (Busemeyer, 1985; Busemeyer, & Townsend, 1993). Here we seek to investigate how new information, past experiences, and deliberation processes affect the accuracy of people's classifications. Accordingly, we explore the effects of three sets of factors on bot-recognition. We focus on 1) new information about the *profiles,* such as their political personas; 2) past experiences and characteristics of the *participants,* such as their extant knowledge of social bots and their party identifications; and 3) the cognitive factors involved in the deliberation *process*, such as attention allocated to decision making.

To examine how these factors affect the ability to differentiate bots from humans, we designed a two-by-two online experiment that included a recognition task involving 20 actual Twitter profiles. The experimental conditions corresponded to test profiles with low vs. high bot/human ambiguity and liberal vs. conservative personas. The profiles were sampled from

followers of US politicians on Twitter. The level of ambiguity was determined using a mix of machine learning and manual coding to select test profiles in the four experimental conditions.

By investigating bots with explicitly partisan personas, we explore whether partisan-motivated reasoning (Redlawsk, 2002) is associated with higher susceptibility to bots among certain individuals and groups. The results demonstrate how the recognition task performance depends on both the partisan personas of the profiles being evaluated and the party affiliations of the evaluators. We explain the discrepancies in performance through the theoretical lenses of motivated-reasoning (Kunder, 1988) and social identity theory (Tajfel, Turner, Austin, & Worchel, 1979). The results have important implications for connecting the two theories (Lelkes & Westwood, 2017) and explaining political polarization on Twitter (Conover et al., 2011).

Social Bots, Political Bots, and Detection

Some social bots appear easily identifiable because they are designed for a single purpose, such as boosting an account's follower or retweet count (Yang et al. 2019). A user can discover the authenticity of the profiles by checking for suspicious profile information, such as automatically generated usernames, unbalanced following-follower ratios, recent profile creation times, missing or incoherent profile descriptions, weirdly distorted profile pictures, etc. (Davis, Varol, Ferrara, Flammini, & Menczer, 2016). In addition, users can check patterns in activity including lack of linguistic diversity, limited number of original tweets, a large number of retweets of the same information, and repetitive interactions with other Twitter users (Davis, et al. 2016).

In comparison with generic social bots, the major damage political bots are designed to do is to impersonate political actors with partisan personas, generating false support and

popularity for a particular party or candidate, and smearing ones' opponents as well as their party (Bessi & Ferrara, 2016; Metaxas & Mustafaraj, 2012; Ratkiewicz et al., 2011). Political bots have also been observed disseminating fake news (Shao et al., 2018a&b), exploiting private user information (Boshmaf et al., 2013), flooding the social media environment with messages of negative sentiment (Stella et al., 2019), and following public figures to create the false impression of popularity (Boshmaf, Muslukhov, Beznosov, & Ripeanu, 2011). Coordinated political bot attacks have been enabled by internet and communication technologies and automation and have been observed to operate on a global scale (Woolley, 2016). Even for bots designed with benign intentions, unsupervised bot-bot interactions can have unexpected consequences (Tsvetkova, García-Gavilanes, Floridi, & Yasseri, 2017). For example, bots that automatically gather and catalog news can become complicit in spreading fake news and conspiracy theories due to a lack of proper human gatekeeping (Lokot & Diakopoulos, 2016).

Contemporary research and resources are mainly concentrated on designing computer-assisted detection systems for social bots in general. Ferrara et al. (2016) summarized the three major types of detection strategies: network-based, crowdsourcing, and feature-based. Networked-based systems are built on the assumption that bot accounts are densely connected to each other, so that finding one leads to finding many. However, advanced bots can easily evade this type of detection by limiting connectivity to other bots. The crowdsourcing strategy relies on the wisdom of the crowd and detects bots using hired workers and trained experts. An early application has demonstrated its high accuracy (Wang, Wang, Zheng, & Zhao, 2014), but the method has low scalability. Moreover, we show in this paper that human biases may affect the reliability of judgments. Feature-based social bot detection systems are more widely adopted. They use supervised Machine Learning (ML) classification to categorize bots based on

individual account characteristics such as the number of friends, number of favorites, number of mentions, account age, etc. Botometer and other supervised ML algorithms use statistical methods (e.g., random forest) or neural networks and have demonstrated satisfying results in identifying individual accounts (Kudugunta & Ferrara, 2018). These algorithms are less effective on novel bots who may appear to be legitimate when inspected individually but can act in coordination (Yang et al., 2019). Ongoing efforts are made to detect these bots by identifying synchrony in the retweeting patterns and highly similar posts among different accounts (Mazza et al., 2019). But this approach faces fundamental challenges in efficiency and lacks flexibility to adapt to different coordination signals. Overall, bot detection is a difficult ML task, particularly because the evolving sophistication of bots generates a constant need for algorithm optimization (or coder training). Using a single strategy has been shown to be insufficient and current bot detection research is moving toward the use of multiple strategies (Ferrara et al., 2016).

While the arms race between computer-assisted detection systems and bot production continues, the role of human perception and the social ramifications of bots, particularly the political ones, have not been completely elucidated yet. As one exception, Bail et al. (2018) showed in an experiment that even users who are aware that they are being exposed to bots with incongruent partisan personas become less trusting of members of the opposite party. In other words, knowingly interacting with political bots exacerbates the existing political polarization. In reality, it is unknown whether users can actually discern political bots from real users.

Investigating the perceptions of political bots is vital for advancing our understanding of digital deception (Hancock, 2007) and automated political manipulation (Woolley, 2016) and for helping us devise proper countermeasures (Gorwa & Guilbeault, 2018). First, by exposing the perceived ambiguity between political bots and real users, we can assess the damage that bots

may cause from a human-centered perspective. Second, uncovering the mechanisms by which users can be deceived and manipulated by political bots may help mitigate such influence via policy interventions and social media literacy campaigns. These mechanisms are likely shared by other online hazards like mis/disinformation (Lazer, et al., 2018; Ginsberg, et al., 2019) and online trolls (Addawood, et al., 2019, Starbird, 2019). Also, while humans have more advanced reasoning capability than ML algorithms, they do not have easy access to many features utilized in ML algorithms and are subject to common cognitive biases. Therefore, understanding the challenges humans face in recognizing political bots can improve ML-assisted bot detection by providing guidance in future coder training.

Profiles, Participants, and the Deliberation Process

Computer-assisted detection systems have not been universally adopted by Twitter users, who are increasingly concerned about the role of bots (Stoking & Sumida, 2018). Differentiating bots from human users relies mostly on individual capacity; it is essentially a deliberative process with considerable uncertainty, especially in an experimental setting. Factors such as past experiences, new information, and cognitive conditions during the deliberation greatly contribute to decision choices (Busemeyer, 1985). In this study, we examine the characteristics of account *profiles*, including their political alignment and bot/human ambiguity; characteristics of *participants*, including their Twitter use behaviors, knowledge of bots, self-reported bot-recognition ability (i.e., self-efficacy), partisan identification, and other demographic characteristics; and attention and perceived uncertainty during the deliberation *process*.

Profiles. Given the broad diversity of political bots, here we focus on two dimensions that may play key roles in affecting human perceptions: the ambiguity between human and bot

accounts, and the partisanship of an account. The ambiguity of bots make detection by simple rules ineffective. For example, machine learning tools to detect social bots extract many features from Twitter metadata—such as the structure of their network of friends and followers—that are not noticeable by a direct observation of the profile (Haustein et al. 2015). While a systematic feature/content analysis of bots is currently lacking, machine learning tools provide consistent metrics to determine the level of ambiguity for bot/human recognition (Yang et al. 2019). We examine whether ambiguous profiles cause more confusion among different users, revealed by lower accuracy in the recognition task (H1).

Researchers have started to notice differentiated activity patterns, effects, and effectiveness of political bots with different partisan personas. Luceri, Deb, Badawy, and Ferrara (2019) analyzed approximately one million Twitter accounts and showed that conservative bots are much more active in terms of the number of tweets; they are twice as active as liberal bots or their conservative human counterparts, and nearly three times as active as liberal humans. Moreover, their research showed that conservative bots are particularly effective compared to liberal bots at establishing following-follower relationships and interactions with humans, e.g., they receive more replies and are retweeted more often. These observations could be explained by a combination of two phenomena: conservative bots being more effective, and/or conservative users being more vulnerable to manipulation by social bots (Grinberg et al. 2019). To explore the first possibility, we hypothesize that participants will have a lower level of accuracy in recognition of conservative profiles, regardless of their party identifications (H2).

Participants. In a recent Pew Research survey, Stoking and Sumida (2018) found that 66 percent of social media users were aware of social bots, 80 percent of them thought bots had negative effects on the public, and 47 percent showed at least some confidence in identifying

bots. Challenging this self-reported confidence, an observational study found that people who were more active on Twitter were more likely to establish mutual following relations with bots (Wagner et al., 2012). To test the observation in a more controlled setting, researchers designed bot accounts with different combinations of realistic features and used them to send private messages to users in an experiment (Wald, Khoshgoftaar, Napolitano, & Sumner, 2013). They found that users were very likely to start following bots, especially when the numbers of friends and followers appeared to be like those of human accounts. These studies emphasize the susceptibility of users. However, they did not acquire any qualitative feedback from the users and therefore it is unknown whether users who followed and interacted with bots also mistook them for real humans.

Here we explore how the activity of Twitter users, their existing knowledge of bots, and their self-estimated ability of recognizing bots contribute to their accuracy in political bot recognition (RQ1). In theory, these factors could affect accuracy in different ways. On the one hand, people learn from past experiences and consequently report more confidence in their competency (Bandura, 2010). This so-called self-efficacy in detecting bots should be positively correlated with better performance in recognition tasks. On the other, theories of egocentric and optimistic biases (Kruger & Dunning, 1999; Kruger & Gilovich, 2004) predict that people tend to overestimate their knowledge and competence and that their self-assessment is either uncorrelated or negatively correlated with their actual ability. Therefore, self-report knowledge and self-assessed competency might not be reliable predictors of performance.

As mentioned in the previous section, Luceri et al. (2019) observed that conservative bots might be more confusing for everyone, but they also noticed that human-bot connections and interactions happened more frequently between conservative bots and conservative users. We

thus want to explore whether conservative users might be more biased when recognizing bots from conservative profiles. In this study, we refer to the theory of motivated reasoning and social identities to examine this kind of bias.

Among other biases, partisan-motivated reasoning is important in the context of political bot recognition. Kunda (1990) defined motivated reasoning as the process whereby "motivation may affect reasoning through reliance on a biased set of cognitive processes—that is, strategies for accessing, constructing, and evaluating beliefs" (p. 480). An extensive literature has documented how salient partisan cues slant people's judgment on politicized issues (Bolsen, Druckman, & Cook, 2014; Kahan, 2012; Nir, 2011; Redlawsk, 2002; Slothuus & De Vreese, 2010). Nir (2011) described motivated reasoning as a tradeoff between two dimensions in political decision making: accuracy seeking and evaluative judgment. She demonstrated that partisan-motivated reasoners reach lower accuracy levels and higher levels of evaluative preferences in judgments.

While motivated reasoning is a commonly applied theory when discussing partisan bias, social identity theory (Tajfel, et al., 1979) offers insights into the possible but different evaluative preferences of partisan-motivated reasoning. According to the theory, the evaluative preferences based on group identities such as partisanships usually take two forms: in-group favoritism and out-group hostility (Brewer, 1999). Interestingly, while both are possible outcomes of partisan-motivated reasoning, they are not necessarily interdependent (Brewer, 1999). Research in partisan-motivated reasoning appears to have produced more evidence for the former (Lelkes & Westwood, 2017). For example, people tend to overestimate public support for their favored candidates (Nir, 2011), are more likely to follow news sources that portray their own parties favorably (Slothuus & De Vreese, 2010), and demonstrate increased support for

favored candidates and groups in the face of negative information (Redlawsk, 2002, see also Kahan 2012).

Lelkes and Westwood (2017) questioned the implicit assumption that in-group favoritism and out-group hostility are interdependent behavioral outcomes in partisan bias research. Their evidence produced a more nuanced understanding of out-group hostility. For example, even those participants who have strong aversive feelings towards the oppositional party were reluctant to *attack* it, but they were open to *avoid* its members if they were asked to form a team (Lelkes & Westwood, 2017). In other words, the non-aggressive behavioral outcome of out-group hostility is much more likely.

In the current context, we focus on the interplay between the partisan affiliations of participants and the partisan personas of test profiles. When people with clear party identifications meet profiles with explicit partisan personas, the setting provides a natural opportunity to test how both aspects of partisan-motivated reasoning—in-group favoritism and out-group hostility—affect judgments. Specifically, we inspect skepticism rather than hostility, because it is not an aggressive behavior and therefore it is more likely to be observed. We hypothesize that people are less inclined to believe that profiles supporting their own parties are impostors (in-group favoritism), and more likely to believe profiles supporting the opposite side to be fake (out-group skepticism) (H3).

Process. While extensive scholarship has been devoted to uncovering cognitive factors in political decision-making *processes*, there is limited evidence of how these factors affect political bot recognition. For example, a workshop report by Wang and colleagues (2014) documented a bot detection system based on crowdsourcing judgments by trained experts and

Mechanical Turk (MTurk) workers. Surprisingly, they found that the duration of the task was not a reliable predictor of accuracy.

While the time participants spend on a task may be a good proxy measure of attention, it is not necessarily the case that more attention means better task performance. The cognitive deliberation process encompasses four steps: detection, recognition, decision, and response (Grimes & Meadowcroft, 1995). Attention is limited, and a fixed amount of cognitive resources can be allocated to the task (Grimes & Meadowcroft, 1995; Lang & Basil, 1998). Usually in simple tasks, a longer reaction time could simply mean that more attention resources are allocated (Lang & Basil, 1998). A shorter time could mean insufficient attention, which leads to poorer performance. However, when subjects are asked to perform more difficult tasks, the longer duration could indicate cognitive-resource deficit due to a higher level of perceived difficulty, which might impede good performance. In other words, time is a nonlinear proxy measure of attention and how it predicts performance is also dependent upon the level of the uncertainty of the task (Busemeyer, 1985).

To further examine Wang's et al. (2014) finding with insight from cognitive psychology, in the present bot detection task we consider the role of two factors in recognition accuracy: time (as a proxy for attention) and perceived uncertainty in the deliberation process (H4). While we explore the potential nonlinear relationship between time and task performance, we are inclined to believe that longer time implies lower accuracy in telling bots from humans, partly because our methodology screens inattentive participants.

Summary of Hypotheses and Research Questions

We formalize and summarize our hypotheses and research question of interest as follows:

H1: Profiles with higher levels of ambiguity will lead to lower accuracy in the recognition task.

H2: Individuals will show lower accuracy in identifying bots among profiles with conservative personas compared to liberal personas.

H3a: People are more likely to confuse humans as bots (i.e., false positives) when the partisan personas are incongruent with their own party identifications.

H3b: People are more likely to confuse bots as humans (i.e., false negatives) when the partisan personas are congruent with their own party identifications.

H4a: People who spend a longer time on deliberation will have a lower level of accuracy in the recognition task.

H4b: People who report a higher level of perceived uncertainty will have a lower level of accuracy in the recognition task.

RQ1: How is political bot recognition accuracy affected by participant characteristics: a) Twitter activity level, b) Twitter behaviors, c) prior knowledge of bots, and d) self-estimated ability of recognizing bots?

Methodology

Procedures

We conducted a two-by-two online within-subject experiment. The two dimensions are ambiguity level and political personas of profiles. The ambiguity level refers to the low or high difficulty of determining whether a social media profile is authentic. A profile's political persona refers to whether the account appears to be a liberal or conservative user. The questionnaire was implemented in Qualtrics. Before the recognition task, all participants were asked about their

Twitter usage behaviors, existing knowledge about social bots, and (self-estimated) ability to recognize bots. They were then asked to participate in a recognition task, after which they answered demographic questions. The recognition task consisted in evaluating 20 Twitter accounts; a preliminary test suggested a higher number was too cumbersome for the participants.

Sample

We recruited only US residents as participants from MTurk. We further screened their data by only including participants who completed all the 20 recognition tasks and spent at least ten minutes on the survey. The cut-off criterion is set by assuming that attentive participants must have spent at least five seconds per question. This rule is rather stringent, as 15.8% of participants did not qualify. After data screening, a total of $N = 656$ subjects were included in the final data analysis. In the final sample, the average age is 35.17 ($SD = 11.62$); 55.79% are female; 75.00% identified as white, 9.90% as African American, 6.40% as Hispanic, 6.40% as Asians or Pacific Islander, and 1.60% as Native American; 22.56% identified as Republican, 40.39% as Democrat, and 37.04% as independent or other. Among all the participants, 91.61% had used Twitter before and 62.19% had at least some knowledge about social bots.

Stimuli

We started from a pool of 28,558 Twitter accounts composed of all followers of US congresspeople from both parties. We used Botometer, a state-of-the-art machine learning algorithm to detect social bots, to determine the *level of ambiguity* of the profiles. Botometer examines over 1,200 features to generate a "bot score" from zero to five for an account: the higher the score, the more likely it is to be automated (Davis et al., 2016; Yang et al., 2019). A very low score indicates little ambiguity in the classification as human, and a very high score

indicates low ambiguity in the classification as bot. A middle score indicates high ambiguity in classification. We selected 1,561 low-ambiguity accounts with bot scores around the two extremes (lower than 0.1 or higher than 4.9), and 785 high-ambiguity accounts with scores around the middle (between 2.48 and 2.52). We conducted a pretest confirming that this ambiguity measure corresponds to different levels of difficulty perceived by participants.

From the selected stimuli, we further randomly sampled profiles to reach a balanced quota of five conservative and five liberal profiles of low ambiguity and five conservative and five liberal profiles of high ambiguity. Figure 1 shows examples of the sampled accounts.

We manually coded the *political personas* of stimuli candidates during the sampling process to reach the intended quota. First, we randomly sampled 20 profiles consisting of two sets of ten from low-ambiguity and high-ambiguity pools respectively and then coded their partisanships. The coding was determined by partisanship-relevant political information in the profile, e.g., words such as "leftist" or "conservative" in the profiles, political codes such as "#MAGA" or "#Resist" in the tweets, and posts about current political affairs or personalities. We then checked the balance of the partisanship in the first 20 profiles, discarded the ones that either did not have visible partisan cues or exceeded the quota, and did additional sampling and coding. Four stimuli candidates were discarded in the first round and six were added (then two discarded). Two of the authors coded partisanship of each account separately and the coding results were in full agreement.

Dependent Variable

Each link to a sampled account was presented to a participant on a separate page with these instructions: "Click on the link, check the profile, and come back to the survey to make a judgment." Participants were then asked, "Do you think the account you just saw is a bot or a

human?" with four reply options: "definitely a bot," "likely a bot," "likely a human," and "definitely a human."

We determined the correct answer, i.e., whether the account was a bot or not, using a hybrid strategy of Botometer scoring, expert coding, and crowdsourcing. First, the sampling process of low-ambiguity profiles based on Botometer scores was used to identify five bots. Following Wang et al (2014), two of the authors served as "expert coders" and coded all cases prior to the experiment. We paid special attention to the highly ambiguous ones, which Botometer struggled to classify. Expert coders were instructed to examine the legitimacy of profile heuristics, e.g., usernames, number of followers/friends, profiles picture, which are commonly noticeable sources of ambiguity (Gorwa, & Guilbeault, 2018) and have a larger weight in ML algorithms (Yang et, 2020). In addition, the expert coders scrutinized timelines and checked if tweets contained natural and diverse language expressions. The results of the two coders were in full agreement with each other and, for the low-ambiguity cases, with Botometer. Third, we compared the results of expert coding with majority answers of participants — essentially trusting their collective intelligence. We sampled n=444 participants with a balanced composition of Republicans, Democrats, and Independents to prevent partisanship bias. The methods above yielded 100% consistent labels.

We then calculated the accuracy of each participant by comparing the answer to each recognition task to the corresponding correct answer, regardless of their perceived uncertainty (i.e., "definitely" or "likely"). We report below on both *accuracy* as a percentage of correct answers and as an odds ratio. Erroneous answers were further recoded into f*alse positive* errors, i.e., humans mislabeled as bots, and f*alse negative* errors, i.e., bots mislabeled as humans.

Independent Variables

Time of deliberation. We recorded the time each participant spent reviewing each profile and making a judgment from the time they clicked on the link to the time they submitted the answer.

Perceived uncertainty. The account classification label was mapped to a binary variable in which one represents "uncertain" (if the account was rated as "likely" rather than "definitely") and zero represents "certain."

Twitter activity and behaviors. We used a five-point Likert scale from "never" to "very frequently" to measure overall Twitter activity level, and eight specific categories of Twitter use behaviors:1) *following celebrities*, 2) *following friends and colleagues*, 3) *checking news*, 4) *checking what's trending other than news*, 5) *posting original tweets*, 6) *retweeting*, 7) *commenting or replying*, and 8) *sending private messages*.

Existing knowledge. Existing knowledge about bots is measured by a five-point scale item. Participants were asked "How much do you think you know about 'social bots'?" The five following options were provided: "A great deal/I'm an expert on this topic," "A lot/I have read quite a lot about them," "Some/I have some knowledge about them," "A little/I have only heard of them," or "None/I have no idea what they are." For participants who chose the last option, the definition of social bots given by Ferrara et al. (2016) was displayed on the next page.

Previous encounters of bots. The frequency of previous encounters with bots is measured by an item asking, "How often do you think you have encountered social bots on social media before?" Participants were provided five options ranging from "never" to "very frequently."

Self-efficacy in recognizing bots. We measured people's self-efficacy by using three 7-point Likert scale items (Cronbach α = .91). Participants were asked to rate from "strongly

disagree" to "strongly agree" the following three statements: 1) "I will recognize most social bots if I encounter them in the future"; 2) "I can succeed at telling social bots apart"; and 3) "When facing social bots that highly resemble regular users, I can still find clues to weed them out."

Demographics. Partisanship was measured using the self-reported categories "Democrat," "Republican," "independent," and "others." "Others" were also provided with open-ended options to specify their political position. We also measured self-reported gender, level of education, and age.

Analytical Framework

To account for unobserved systematic differences between individuals, we used two-level mixed-effect logistic regression models. The unit of analysis at level one is per participant per recognition ($n = 656 * 20$) and the unit of analysis at level two is the individual participant ($N = 656$). We ran four models: a null model with only the intercepts fitted, model 1 with deliberation process, model 2 adding characteristics of profiles, and model 3 adding characteristics of participants. Due to the large number of independent variables, we focus on results regarding the main variables of interest (namely, the partisan persona of profiles and the partisanship of participants) and other significant factors related to Twitter use, knowledge, and bot recognition self-efficacy.

Results

Moderate Capability and Perceived Difficulty

On average, the participants demonstrated a moderate capability of identifying bots with an accuracy of 71% ($SD = 12\%$) on the 20 profiles. It is statistically significantly higher than chance ($t = 43.22$, $df = 655$, $p < .001$). As manipulation checks, we examined whether

participants perceive some profiles to be more difficult to make a judgment by comparing perceived uncertainty and time of deliberation between low- and high-ambiguity profiles. The results showed that highly ambiguous profiles were classified with higher perceived uncertainty (MD = 5%, t = −3.87, df = 1,306.5, p < .001), and participants took significantly longer to make decisions about them (MD = 4 seconds, t = −3.89, df = 1,193.3, p < .01). Other descriptive statistics are summarized in Table 1.

The results of H1 and H4 further corroborate that participants were cognizant of different levels of human/bot ambiguity in different profiles. H1 predicts that the Botometer-based ambiguity measure corresponds to the level of accuracy in participant performance. The average accuracy rate for differentiating bots and humans is approximately 69% (SD = 17%) among highly ambiguous cases and 73% (SD = 16%) among low-ambiguity cases. Although the numeric difference is small, a Welch t-test shows a statistically significant difference (t = 4.27, df = 1,306.5, p < .01). Controlling for other factors in model 3, the ambiguity level accounts for approximately a 2% accuracy drop (β = −.13, p < .01). H1 therefore is supported.

H4 concerns the deliberation process: both longer periods of deliberation and higher perceived uncertainty of profiles predict lower recognition accuracy. Results of model 3 show that a one-minute increase in deliberation contributes to an accuracy drop of less than one percent (β = -.08, p < .05), and perceived uncertainty contributes to a 14% accuracy drop (β = −.92, p < .01).

To test the potential non-linear relationship between time of deliberation and recognition accuracy, we added a second-order term of the time variable. After including the second order-term, the first order time variable contributed to a greater accuracy drop of 2.5% (β = −.21, p < .01), while the second-order term showed statistical significance with positive but minimal

effects (β = .02, p < .05). However, accuracy monotonically decreases with time even in the polynomial regression, confirming that our experimental conditions exclude inattentive participants. We also tested with the order of profiles as a third factor in the recognition process, to account for potential learning effects. We found no significant increase in accuracy during the task that could be attributed to learning.

Asymmetrical Perceptions and Partisan-Motivated Reasoning

We investigate the effects of partisan personas of profiles and party affiliations of participants and their interplay in H2 and H3. H2 contends that conservative profiles are more effective in creating confusion. The results support this hypothesis as well: participants, regardless of their party affiliations, have significantly lower accuracy rates (t = −4.58, df = 1,301, p < .01) when differentiating bots and humans among profiles with conservative personas (M = 69%, SD = 15%) than among ones with liberal personas (M = 73%, SD = 16%).

To interpret the higher error rate in identifying conservative versus liberal profiles, we further explore the two types of mistakes: *false positive* and *false negative* errors. We find that conservative bots are significantly more likely to be misidentified than liberal bots (MD = 9%, t = 14.15, df = 1,203.1, p < .01), whereas conservative humans are significantly less likely to be mislabeled as bots than liberal humans (MD = −5%, t = −8.87, df = 1,162.4, p < .01). Therefore, the confusion can be attributed to missing conservative bots.

A focus of this study is partisan-motivated reasoning, i.e., how participant perceptions depend on whether the political personas of accounts are congruent with their own party identification. H3 predicts that participants are more likely to a) mislabel incongruent human accounts as bots and b) miss bots with congruent political personas. Only Democrats and

Republicans are included in this analysis (n = 413). First, we recoded participant-profile pairs into two groups: Republican-conservative profiles and Democrat-liberal profiles as the congruent group, and Republican-liberal profiles and Democrat-conservative profiles as the incongruent group. To our surprise, we did not find support for H3. The average false positive error rate is 12% (SD = 12%) for congruent cases and 13% (SD = 12%) for incongruent ones. The average false negative rate is 16% (SD = 14%) for congruent cases and 18% (SD = 13%) for incongruent ones. The differences are not significant ($p > .05$).

We then refined our analysis of H3 by contrasting congruent versus incongruent pairs within conservative and liberal profiles separately. In other words, as illustrated in Figure 2, we compared Republican-conservative profile pairs against Democrat-conservative profile pairs and Democrat-liberal profile pairs against Republican-liberal profile pairs. The results show that H3a and H3b are actually supported for the case of conservative profiles: Democrats are significantly more likely than Republicans to mislabel conservative humans as bots, i.e., out-group skepticism (MD = 2%, t = 2.04, df = 304.42, $p < .05$) and Republicans are significantly more likely than Democrats to miss conservative bots, i.e., in-group favoritism (MD = 5%, t = 3.15, df = 288.21, $p < .01$). However, we do not observe the same pattern with statistical significance for liberal profiles (both $p > .05$). Therefore, the hypothesized partisan-motivated reasoning is also asymmetrical in the sense that it is only supported when judging conservative profiles.

Predicting Performance in Bot Recognition

The null model yields a predicted baseline accuracy of 72% (estimated fixed intercept .92, $p < .001$ and random intercept .13 at level two). After including incorporating features of profiles, Twitter use habits and demographic characteristics of participants, and the

deliberation process, the predicted baseline accuracy of model 3 increases to 88%. Overall, the model fit shows significant improvement (i.e., the null model: AIC = 15,744, BIC = 15,760; model 3: AIC = 13,969, BIC = 14,051). The intra-class correlation (ICC) shows that the unobserved individual differences explain only 4% of the total variance in the null model. However, incorporating the participant characteristics does not reduce much of the unobserved individual variance in model 3 compared with models 1 and 2: the ICC remains the same. Table 2 shows model coefficients and goodness-of-fit indices allowing us to test our hypotheses.

To our surprise, among 14 variables included in testing the prediction of Twitter use habits and demographic characteristics of participants (*RQ1*), only three factors are significant and therefore preserved in model 3. They are frequency of checking news ($\beta = .06$, $p < .01$), frequency of sending private messages ($\beta = -.07$, $p < .01$), and age ($\beta = -.05$, $p < .05$). Each of these factors contributes only less than one percent to the accuracy. It is particularly worth noting that neither extant knowledge of social bots nor self-efficacy of recognizing bots are significant predictors (both $p > .05$). However, controlling for these factors, H2 and H3 remain supported: significantly more errors are made when participants are Republicans ($\beta = -.21$, $p <. 01$) and profiles have conservative personas ($\beta = -.22$, $p < .01$), with accuracy drops of 2.5% and 2.6% respectively. Due to the noted asymmetry in H3 testing, we also included in the model an interaction term between political personas of profiles and party identification of participants. The results yield no statistical significance.

Discussion

Social media platforms are being manipulated through malicious automated algorithms to sway public opinion for political gains (Ferrara, 2016; Yang et al., 2019). This is one of the first studies to explore the human capability of recognizing political bots. We designed a recognition

task in an online experiment that explores the role of *profiles*, *participants*, and *process* in bot detection accuracy.

The results of the experiment provide new evidence of individual (in)ability to differentiate low-credibility actors on social media. We found that humans perform well in classifying low-ambiguity accounts but have difficulty in differentiating ambiguous ones. Yet, human participants manage to perform significantly better than chance even in ambiguous cases. But partisan information, interacting with partisan minds, impairs the capability.

By examining how the recognition accuracy is affected by the political personas of profiles and the political affiliations of participants, we found an asymmetrical perception of partisan political bots. We experimentally confirm the observational findings of Luceri et al. (2019) that conservative bots are more effective than liberal bots at establishing interactions with humans. Our results also suggest two mechanisms to explain this effect: conservative bot profiles are harder to detect, and conservative humans are more vulnerable. The susceptibility of Republican participants to bots is also consistent with the finding that conservative voters were more likely to read and share false news during the 2016 U.S. election (Grinberg et al., 2019).

What is more interesting in our results is that there are two main reasons why conservative profiles are confusing. First, conservative participants are more likely to be deceived by conservative bots. Second, liberal participants are more likely to mislabel conservative humans as bots. These effects are both likely associated with partisan-motivated reasoning, i.e., participants with different party identity rely on partisan cues of the profile instead of other information to make judgment (Redlawsk, 2002). These results also confirm that the two sides of partisan-motivated reasoning —in-group favoritism and out-group skepticism— are not necessarily both observed in the same group (Brewer, 1999; Lelkes & Westwood, 2017).

The asymmetry between Republicans who display stronger in-group favoritism and Democrats who exhibit greater out-group skepticism could reflect the contemporary political zeitgeist, where the incumbent president is a Republican who is highly active on Twitter (Wells et al, 2016). Therefore, Republicans and Democrats may be more likely to be in "protect" and "attack" modes, respectively, toward highly active Twitter accounts.

The results about other characteristics of participants indicate that previous experience with social bots as well as levels and types of Twitter activity have little effect on bot detection accuracy. Our explanation leans toward a possible egocentric/optimistic bias (Kruger & Gilovich, 2004): self-assessment of competence in recognizing bots is unreliable and likely inflated. The results also show that older participants appear to be more vulnerable. Although the effect size is small, the age disparity is consistent with observations by Grinberg et al. (2019), who found that older adults were more likely to engage with misinformation during the 2016 U.S. presidential election.

Finally, time of deliberation and perceived uncertainty negatively contribute to the correct identification of bots. This may be due in part to the minimum duration threshold that we imposed in data cleaning to limit errors due to lack of attention. Although there was no maximum time limit, the longest deliberation time for a single profile classification was around six minutes. Therefore, it is unlikely that subjects employing more time were distracted. The negative correlation between time of deliberation and performance suggests that participants paid sufficient attention but were unable to make decisions with confidence about their correctness. In other words, greater efforts do not yield better results. The observed 70% accuracy for highly ambiguous cases may reflect a limit in the human capability of detecting political bots.

Limitations and Future Research

As one of the first studies investigating the human perception of political bots, our experiment is not without caveats. First, the sample of participants is not nationally representative, and in particular overrepresents liberals. We used MTurk for at least two reasons: participants have similar demographics to Twitter users, and the platform allows for fast data collection. Because we used real Twitter profiles, it was necessary to finish the data collection within a short period of time so that every participant saw similar profiles. On the other hand, we acknowledge concerns about the quality of data collected from MTurk (Levay, Freese, & Druckman, 2016).

By using real Twitter profiles, our experiment aimed for ecological validity at the cost of slightly weakened experimental control. Future research could build experimental environments in which researchers have full control over the profiles. Furthermore, because the stimuli resulted from purposive sampling, the findings related to profile features may not be generalizable to all political bots. Considering the complex and diverse design of bots, future research will benefit from large-scale content analysis, where reliability of the coding is tested on an independent sample. Experiment research could use a larger random sample of stimuli to account for systematic differences among profiles. In particular, the deceptive nature of conservative profiles needs to be further explored in studies with representative samples of partisan bots. Finally, to further test partisan differences in motivated reasoning, it would be of interest to replicate the study when the president is from the Democratic Party.

Combating the Culture of Distrust

Our study demonstrates that users have moderate capability to identify political bots, but such a capability is also limited due to cognitive biases. In particular, bots with explicit political personas activate the partisan bias of users. Machine learning algorithms to detect social bots provide one of the main countermeasures to malicious manipulation of social media. While adopting third-party bot detection methods is still advisable, our findings also suggest possible bias in human-annotated data used for training these machine learning models. This also calls for a careful consideration of algorithmic bias in future development of artificial intelligence tools for political bot detection.

The very existence of political bots is fueling the culture of distrust, wherein not only false accounts are believed to be authentic, but also authentic accounts are dismissed as fake. Like recently proposed countermeasures for fake news (Bakir & McStay, 2018), solutions to the problem of political bots require collective efforts by multiple stakeholders. These may include laws and policies for dealing with deceptive bots (Lamo & Calo, 2019; Gorwa & Guilbeault, 2018), tracing money sources and massive deployment in the early stage, and devising social media literacy campaigns (Yang et al., 2019). Social media users may need to adopt hybrid strategies in dealing with bots: checking profiles directly, using third-party bot detection tools, consulting multiple sources if possible, and becoming aware of their own biases.

References

Addawood, A., Badawy, A., Lerman, K., & Ferrara, E. (2019, July). Linguistic cues to deception: Identifying political trolls on social media. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 13, No. 01, pp. 15-25).

Badawy, A., Ferrara, E., & Lerman, K. (2018). Analyzing the digital traces of political manipulation: The 2016 Russian interference twitter campaign. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (pp. 258–265). doi: 10.1109/ASONAM.2018.8508646

Bail, C. A., Argyle, L. P., Brown, T. W., Bumpus, J. P., Chen, H., Hunzaker, M. F., ... & Volfovsky, A. (2018). Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences*, 115(37), 9216-9221.

Bakir, V., & McStay, A. (2018). Fake news and the economy of emotions: Problems, causes, solutions. *Digital journalism*, 6(2), 154-175.

Bandura, A. (2010). Self-efficacy. *The Corsini Encyclopedia of Psychology*, 1–3.

Brewer, M. B. (1999). The psychology of prejudice: Ingroup love and outgroup hate?. *Journal of social issues*, 55(3), 429-444.

Bessi, A., & Ferrara, E. (2016). Social bots distort the 2016 us presidential election online discussion. *First Monday*, *21*(11). Retrieved from https://firstmonday.org/ojs/index.php/fm/article/view/7090 doi: 10.5210/fm.v21i11.7090

Bolsen, T., Druckman, J. N., & Cook, F. L. (2014). The influence of partisan motivated reasoning on public opinion. *Political Behavior*, *36*(2), 235–262. doi: 10.1007/s11109-013-9238-0

Boshmaf, Y., Muslukhov, I., Beznosov, K., & Ripeanu, M. (2011). The socialbot network: when bots socialize for fame and money. In *Proceedings of the 27th Annual Computer Security Applications Conference* (pp. 93–102).

Boshmaf, Y., Muslukhov, I., Beznosov, K., & Ripeanu, M. (2013). Design and analysis of a social botnet. *Computer Networks*, *57*(2), 556–578. doi: https://doi.org/10.1016/j.comnet.2012.06.006

Burkner, P.-C. (2018). Advanced bayesian multilevel modeling with the r package brms. *The R Journal*, *10*(1), 395–411. doi: 10.32614/RJ-2018-017

Busemeyer, J. R. (1985). Decision making under uncertainty: a comparison of simple scalability, fixed-sample, and sequential-sampling models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*(3), 538. doi: 10.1037/0033-295x.100.3.432

Busemeyer, J. R., & Townsend, J. T. (1993). Decision field theory: a dynamic-cognitive approach to decision making in an uncertain environment. *Psychological review*, 100(3), 432. doi: 10.1037//0033-295x.100.3.432

Conover, M. D., Ratkiewicz, J., Francisco, M., Gonçalves, B., Menczer, F., & Flammini, A. (2011). Political polarization on twitter. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media.*

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* (pp. 273–274).

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

Gardiner, J. M., Ramponi, C., & Richardson-Klavehn, A. (2002). Recognition memory and decision processes: A meta-analysis of remember, know, and guess responses. *Memory*, 10(2), 83-98.

Gorwa, R., & Guilbeault, D. (2018). Unpacking the social media bot: A typology to guide research and policy. Policy & Internet.

Grimes, T., & Meadowcroft, J. (1995). Attention to television and some methods for its measurement. *Annals of the International Communication Association*, 18(1), 133-161. doi: 10.1080/23808985.1995.11678910

Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., & Lazer, D. (2019). Fake news on Twitter during the 2016 US presidential election. *Science,* 363(6425), 374-378.

Hart, P. S., & Nisbet, E. C. (2012). Boomerang effects in science communication: How motivated reasoning and identity cues amplify opinion polarization about climate mitigation policies. *Communication Research*, 39(6), 701-723.

Hancock, J. T. (2007). Digital deception. *Oxford handbook of internet psychology*, 289-301.

Kahan, D. M. (2012). Ideology, motivated reasoning, and cognitive reflection: An experimental study. *Judgment and Decision Making*, *8*, 407–24. doi: 10.2139/ssrn.2182588

Kiley, J. (2017, Oct). In Polarized era, fewer Americans hold a mix of conservative and liberal views. *Pew Research Center*. Retrieved from https://www.pewresearch.org/fact-tank/2017/10/23/in-polarized-era-fewer-americans-h

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, 77(6), 1121. doi: 10.1037/0022-3514.77.6.1121

Kruger, J., & Gilovich, T. (2004). Actions, intentions, and self-assessment: The road to self-enhancement is paved with good intentions. *Personality and Social Psychology Bulletin*, *30*(3), 328–339. doi: 10.1177/0146167203259932

Kudugunta, S., & Ferrara, E. (2018). Deep neural networks for bot detection. *Information Sciences*, 467, 312-322.

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

Lamo, M., & Calo, R. (2019). Regulating Bot Speech. *UCLA L. Rev.*, 66, 988.

Lang, A. & Basil, M. D. (1998). Attention, resource allocation, and communication research: What do secondary task reaction times measure, anyway?. *Annals of the International Communication Association*, 21(1), 443-458. doi: 10.1080/23808985.1998.11678957

Lelkes, Y., & Westwood, S. J. (2017). The limits of partisan prejudice. *The Journal of Politics*, 79(2), 485-501.

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

Lokot, T., & Diakopoulos, N. (2016). News Bots: Automating news and information dissemination on Twitter. *Digital Journalism*, *4*(6), 682-699.

Luceri, L., Deb, A., Badawy, A., & Ferrara, E. (2019, May). Red bots do it better: Comparative analysis of social bot partisan behavior. In *Companion Proceedings of The 2019 World Wide Web Conference* (pp. 1007-1012).

Mazza, M., Cresci, S., Avvenuti, M., Quattrociocchi, W., & Tesconi, M. (2019, June). RTbust: exploiting temporal patterns for botnet detection on twitter. In *Proceedings of the 10th ACM Conference on Web Science* (pp. 183-192).

Metaxas, P. T., & Mustafaraj, E. (2012). Social media and the elections. *Science*, *338*(6106), 472–473. doi: 10.1126/science.1230456

Nir, L. (2011). Motivated reasoning and public opinion perception. *Public Opinion Quarterly*, *75*(3), 504–532. doi: 10.1093/poq/nfq076

Ratkiewicz, J., Conover, M. D., Meiss, M., Gonçalves, B., Flammini, A., & Menczer, F. M. (2011). Detecting and tracking political abuse in social media. In *Fifth International AAAI Conference on Weblogs and Social Media.*

Redlawsk, D. P. (2002). Hot cognition or cool consideration? testing the effects of motivated reasoning on political decision making. *The Journal of Politics*, *64*(4), 1021–1044. doi: 10.1111/1468-2508.00161

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

Shao, C., Hui, P. M., Wang, L., Jiang, X., Flammini, A., Menczer, F., & Ciampaglia, G. L. (2018b). Anatomy of an online misinformation network. *PloS One,* 13(4), e0196087.

Slothuus, R., & De Vreese, C. H. (2010). Political parties, motivated reasoning, and issue framing effects. *The Journal of Politics*, *72*(3), 630–645. doi: 10.1111/pops.12164

Starbird, K. (2019). Disinformation's spread: bots, trolls and all of us. *Nature*, 571(7766), 449.

Stella, M., Ferrara, E., & De Domenico, M. (2018). Bots increase exposure to negative and inflammatory content in online social systems. *Proceedings of the National Academy of Sciences*, 115(49), 12435-12440.

Stoking, G., & Sumida, N. (2018, Oct 15). Social media bots draw public's attention and concern. *Pew Research Center*. Retrieved from https://www.journalism.org/2018/10/15/social-media-bots-draw-publics-attention-and-

Tajfel, H., Turner, J. C., Austin, W. G., & Worchel, S. (1979). An integrative theory of intergroup conflict. *Organizational identity: A reader*, 56-65.

Tsvetkova, M., García-Gavilanes, R., Floridi, L., & Yasseri, T. (2017). Even good bots fight: The case of wikipedia. *PloS one*, *12*(2), doi: 10.1371/journal.pone.0171774

Wagner, C., Mitter, S., Körner, C., & Strohmaier, M. (2012). When social bots attack: Modeling susceptibility of users in online social networks. In *The Second Workshop on Making Sense of Microposts* (pp. 41–48).

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). doi: 10.1109/iri.2013.6642447

Wang, G., Wang, T., Zheng, H., & Zhao, B. Y. (2014). Man vs. machine: Practical adversarial detection of malicious crowdsourcing workers. In the *Proceedings of the 23rd USENIX Security Symposium*. *14)* (pp. 239–254).

Woolley, S. C. (2016). Automating power: Social bot interference in global politics. *First Monday*.

Wells, C., Shah, D. V., Pevehouse, J. C., Yang, J., Pelled, A., Boehm, F., ... & Schmidt, J. L. (2016). How Trump drove coverage to the nomination: Hybrid media campaigning. *Political Communication*, 33(4), 669-676.

Yang, K.-C., Varol, O., Davis, C. A., Ferrara, E., Flammini, A., & Menczer, F. (2019). Arming the public with artificial intelligence to counter social bots. *Human Behavior and Emerging Technologies*, *1*(1), 48–61. doi: 10.1002/hbe2.115

Yang, K. C., Varol, O., Hui, P. M., & Menczer, F. (2020, April). Scalable and generalizable social bot detection through data selection. *In Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 34, No. 01, pp. 1096-1103).

Table 1: Descriptive data (N = 656)

Dependent variables:	Mean	SD
Overall accuracy	71%	12%
False Positive	12%	9%
False Negative	17%	10%
Deliberation Process		
Time of deliberation (seconds)	23.83	17.39
Perceived uncertainty	67%	23%
Characteristics of participants		
Activity Level (1–5)	3.52	1.17
Twitter Behavior (1–5)		
1) Following celebrity	3.16	1.20
2) Following friends and colleagues	3.48	1.21
3) Checking news	3.54	1.17
4) Checking what's trendy other than news	3.27	1.16
5) Posting original tweets	2.73	1.18
6) Sharing and retweeting	2.91	1.21
7) Commenting and replying	2.83	1.18
8) Sending private message	2.35	1.22
Knowledge of bots (1–5)	2.71	0.87
Frequency of encountering bots (1–5)	3.01	1.01
Self-assessed ability of recognizing bots (1–7)	4.79	1.18

Table 2: Model fitting: Predicting odds ratios of correct recognitions (1 = correct)

	Model 1		**Model 2**		**Model 3**	
Fixed Effects	*Odds Ratios*	*Estimate (SE)*	*Odds Ratios*	*Estimate (SE)*	*Odds Ratios*	*Estimate (SE)*
(Intercept)	5.03 ***	1.62 (.05)	6.02 ***	1.79 (.05)	7.24 ***	1.98 (.14)
Time (unit = min)	.91 **	-.09 (.04)	.92 *	-.08 (.04)	.92 *	-.08 (.04)
Perceived uncertainty	.40 ***	-.91 (.05)	.40 ***	-.91 (.05)	.40 ***	-.92 (.05)
High-ambiguity[a]			.87 ***	-.14 (.04)	.87 **	-.13 (.04)
Conservative[b]			.81 ***	-.22 (.04)	.80 ***	-.22 (.04)
Republicans[c]					.81 **	-.21 (.07)
Democrats[c]					.93	-.07 (.06)
Checking news					1.07 **	.06 (.02)
Private message					.93 ***	-.07 (.02)
Age (unit = SD)					.95 *	-.05 (.03)
Random Effects						
Variance (level 2)	.15		.15		.13	
ICC	.04		.04		.04	
Marginal R^2 / Conditional R^2	.05 / .09		.06 / .10		.07 / .10	
AIC	15343		15304		13970	
BIC	15373		15349		14051	

Notes: Reference group a) low-ambiguity, b) liberal personas, c) Independents. * *p<0.05* ** *p<0.01* *** *p<0.001*

Figure 1: Examples of low-ambiguity liberal (top left), low-ambiguity conservative (top right), high-ambiguity liberal (bottom left) and high-ambiguity conservative (bottom right) profiles used as stimuli. Identifiable information is blurred.

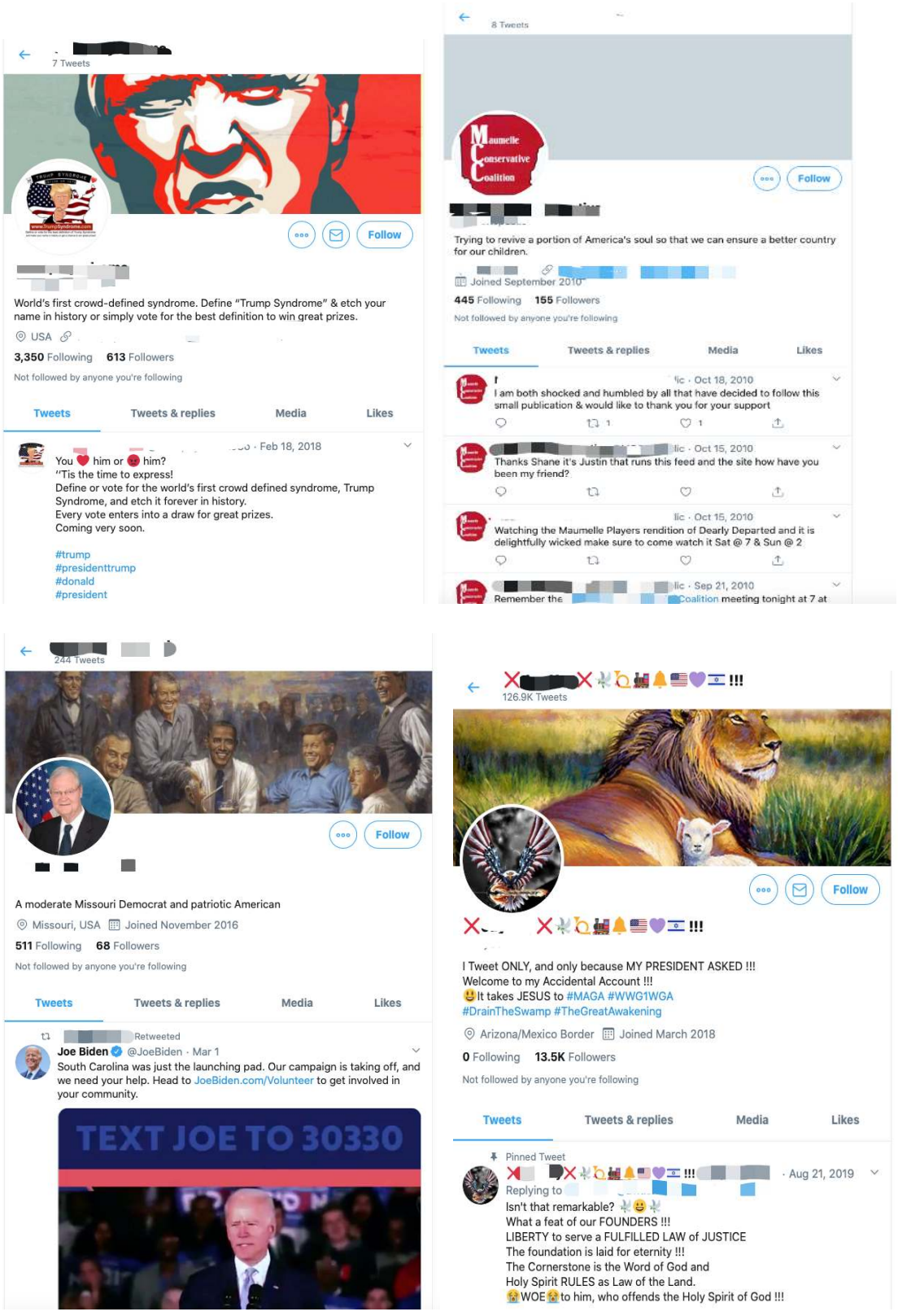

Figure 2: Partisan bias for profiles with (top) conservative personas and (bottom) liberal personas.