"I Have No Idea What a Social Bot Is": On Users' Perceptions of Social Bots and Ability to Detect Them

Daniel Kats
daniel.kats@nortonlifelock.com
NortonLifeLock Research Group
Culver City, CA, USA

Mahmood Sharif
mahmoods@cs.tau.ac.il
Tel Aviv University
Tel Aviv, Israel

ABSTRACT

Social bots—software agents controlling accounts on online social networks (OSNs)—have been employed for various malicious purposes, including spreading disinformation and scams. Understanding user perceptions of bots and ability to distinguish them from other accounts can inform mitigations. To this end, we conducted an online study with 297 users of seven OSNs to explore their mental models of bots and evaluate their ability to classify bots and non-bots correctly. We found that while some participants were aware of bots' primary characteristics, others provided abstract descriptions or confused bots with other phenomena. Participants also struggled to classify accounts correctly (e.g., misclassifying >50% of accounts) and were more likely to misclassify bots than non-bots. Furthermore, we observed that perceptions of bots had a significant effect on participants' classification accuracy. For example, participants with abstract perceptions of bots were more likely to misclassify. Informed by our findings, we discuss directions for developing user-centered interventions against bots.

ACM Reference Format:
Daniel Kats and Mahmood Sharif. 2022. "I Have No Idea What a Social Bot Is": On Users' Perceptions of Social Bots and Ability to Detect Them. In *Proceedings of the 10th International Conference on Human-Agent Interaction (HAI '22), December 5–8, 2022, Christchurch, New Zealand.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3527188.3561928

1 INTRODUCTION

Social bots have gained increased attention in recent years. These are typically automated or semi-automated accounts on online social networks (OSNs) controlled by computer software while exhibiting anthropomorphic behavior. While bots are sometimes used for benign aims such as assisting users in reading long content [29], they are often deployed for malicious purposes, such as disseminating malware [20], participating in spam and phishing campaigns [49], spreading disinformation [33], popularizing conspiracy theories [22], and polarizing online conversations [52].

In return, researchers have invested significant efforts to understand the behavioral properties of bots, and the mechanisms that drive their flourishing ecosystem [1, 27, 36]. Guided by this understanding, researchers developed numerous bot-detection technologies to help moderate bots and counter their abuse [11, 17, 58]. These technologies, sometimes further supported by human experts, helped OSNs suspend many malicious bot campaigns [55].

Despite these efforts, malicious bots remain widely active on popular OSNs. Varol et al. estimated that 9%–15% of active Twitter accounts are bots [54]. More recently, whistleblower reported that OSNs took no actions against malicious bots despite being aware of their activity [9, 18]. Thus, as the burden of detecting bots falls on OSN users, there remains a need for improved interventions, technological and otherwise, to help them identify bots and mitigate attacks. Prior work mainly considered technical issues related to bot detection, and, except for a few efforts [5, 6, 35], mostly ignored human aspects. Nonetheless, a deeper understanding of users (how they perceive bots, the types of users who need help, ...) can help inform the development of effective user-centered interventions.

To this end, we conducted an online study with 297 participants to *1)* gauge their perceptions of bots and whether they believe bots can influence their behavior; *2)* assess whether they can classify bots and non-bots correctly and quantified how different factors impact their accuracy; and *3)* measure their satisfaction with different OSNs and how they believe OSNs should handle bots (see §3). Among others (§4), we found that participant perceptions of bots captured real aspects of bot operation (e.g., automation) and goals (e.g., artificially boosting content popularity). Nonetheless, some participants confused bots with other phenomena (e.g., online tracking and personalization) or had abstract perceptions bots. A substantial portion (52.53%) of participants believed that bots do not influence them personally, including because they believed they can accurately detect bots. Yet, participants could not correctly classify all accounts we showed them (31.70% avg. error). Our analysis indicated that multiple factors impact user likelihood to classify accounts correctly, including their perceptions of bots (e.g., users with abstract perceptions were more likely to misclassify) and the types of accounts (e.g., fake-follower bots were misclassified more than other accounts). Lastly, participants had varied satisfaction levels with how OSNs governed bots—they were particularly unsatisfied with OSNs they thought had significant presence of bots—and had diverse opinions on how they should do so. Based on our findings, we make recommendations to improve interventions (§5).

2 RELATED WORK

This section reviews common definitions of bots, and summarizes work on bot detection and human-bot interaction.

2.1 Definitions of Social Bots

Finding a standard definition of *social bots* is surprisingly challenging. Grimme et al. tackled this difficulty and concluded that most definitions attribute a level of automation to bots, but may also include motivations (e.g., maliciousness), capabilities (e.g., content generation), and additional specificity (e.g., mimicking humans) [27]. In our meta-analysis of 15 recent papers [3, 4, 11, 13, 17, 19, 21, 23, 30, 35, 38, 53, 54, 57, 58], researchers often mentioned human imitation, automated interaction with other accounts, and automated content generation in their definitions. Definitions also mentioned the term "fake" and platform abuse. In this work, we compare users' mental

models of bots with researcher definitions, characterizing ways it which they agree and disagree (§4.1).

Definitions generally identify that bots can be benign or malicious [51]. Benign bots can provide entertainment or assist humans in various means, such as aggregating news, moderating forums, and supporting customers [23, 29, 34, 35]. By contrast, malicious bots can be used in different ways, including artificially boosting influence on OSNs via "fake followers" [15], influencing people's opinions and increasing polarization [53], spreading malware and disinformation [20, 33], and participating in spam and phishing [60].

2.2 Social Bot Detection

Substantial work has gone into developing bot-detection algorithms on OSNs, and especially Twitter, due to the public visibility of most of its content and a well-documented API [11, 17, 19, 38, 58]. There have also been efforts to detect bots on other OSNs, including Facebook [46] and Instagram [48]. Most bot-detection algorithms rely on machine-learning (ML) algorithms [4, 14, 19, 58]. Typically, ML models are trained to distinguish bots from non-bots using a collection of features characterizing account metadata (e.g., follower-to-following), post and profile contents (e.g., hashtag prevalence), network structures, and account behavior over time. We contrast how user bot-detection strategies compare with ML algorithms, and suggest ways to help users detect bots accurately, including with the help of ML (§4.2 and §5.2).

There have been attempts to equip users with bot-detection tools. BotOMeter is a web app users can query with Twitter handles to be classified [58]. BotSight, a tool available as a browser extension or mobile app, inlines Twitter posts and profiles with indicators denoting the likelihood accounts are bots. While beneficial, these tools have small user bases (e.g., ∼500 daily BotOMeter visitors [57]), and are unlikely to be adopted by the users expected to benefit from them the most (specifically, those unfamiliar with bots §5.2).

2.3 User-Bot Interaction

Prior work discovered that user susceptibility to manipulation by political and spam bots is varied. Badawy et al. found that holding conservative political views was a primary predictor for whether users would spread disinformation [6], while Luceri et al. attributed conservative bots' effectiveness at spreading disinformation to network centrality [36]. Redmiles et al. characterized users that click on spam and found that users less active on the Internet were more susceptible than others [44]. By contrast, we seeks to characterize OSN users' mental models of bots and their ability to differentiate between bots and non-bots. Moreover, we do not limit ourselves to political or spam bots, but consider diverse account types.

Researchers have investigated user ability to detect bots and disinformation. Alarifi et al. asked nine study participants to label accounts as bots and non-bots and measured 96% accuracy after classifying accounts by a vote over participant labels [2]. Similarly, Cresci et al. asked study participants to label accounts which they later classified by a vote [16]. Consistently with our findings, many of their participants thought that bots that belong to a particular family (specifically, spam bots) were non-bots. Unlike ours, however, prior studies did not quantify individuals' detection accuracy, considered limited account types, and did not explore participant strategies used to classify accounts. Freeman concluded that average users cannot accurately detect fraudulent activity, but did not explore user perceptions and detection strategies [24]. Geeng et al. reported that their study participants—unaware of being tested—were unlikely to investigate disinformation planted into OSN feeds [25]. They attributed some participants' lack of investigation to over-confidence in their ability to spot misinformation. Our participants exhibited similar behavior (§4.1). Finally, Appling and Briscoe asked their study participants to select account features that were most associated with bots [5]. Participants denoted that links in the profile and large numbers of posts were most indicative of bots. We study the features users rely on in further detail (§4.2).

Prior work has touched on mental models of bots. Chaves et al. and Zamora studied user expectations from chatbots [10, 59]. They found users were dissatisfied with bots' ability to understand questions and produce quality responses. Clément et al. analyzed posts on Wikipedia discussion pages to study reactions to bots [12]. They grouped bots into two ideotypes: "servant" and "policing" bots. User reactions to bots varied depending on ideotypes, with policing bots eliciting more polarized responses. Wessel et al. examined bots as assistants during software development [56]. They surveyed developers to gauge their perceptions of bots and found bots were not sufficiently advanced to impact the development process. Gero et al. asked participants to play a game against an automated agent and found that those who won had more accurate models of the agent's capabilities [26]. Finally, Long et al. studied Reddit to explore user perceptions of benign bots [35]. They found that perceptions varied with technical ability: while most users understood the basic bot functions, non-technical users often had unrealistic expectations from bots. Unlike prior work, we study user perceptions of both benign and malicious bots, and do not restrict ourselves to a single OSN or automated agent when exploring perceptions. Moreover, we discover new concepts that OSN users may identify with bots, sometimes erroneously (e.g., online tracking).

3 METHODOLOGY

We now present our study design and analysis methods, followed by the study's participants and limitations.

3.1 Study Design

We designed an online (crowdsourced) study to explore mental models of bots, measure user ability to detect them, and assess user preferences for how to moderate bots. The study began by asking participants about general OSN usage habits. Particularly, we inquired about whether participants have accounts on any of the following seven OSNs: Facebook, Instagram, LinkedIn, Reddit, TikTok, Twitter, and YouTube. Then, for each OSN a participant has an account on, we asked about how often they visit the OSN, how much time they spend on it, and what content types they create or consume. We selected these specific seven OSNs because they represent a variety of popular OSNs known to host bot activity and differ in several interesting aspects, including their purpose, user demographics, and the modes of interactions between users. For example, TikTok is a six-year-old OSN primarily used by young users for entertainment. In contrast, LinkedIn is a 17-years-old OSN used by working professionals to share professional content.

Next, we elicited perceptions of bots, and participants' belief on whether bots influence OSN users. Specifically, we asked whether participants were familiar with bots and instructed them to provide concise definitions of bots and their goals in free-form text. If participants had not encountered the term bot before, we encouraged them to best guess the definitions and goals. Moreover, we asked how common participants believe bots are on OSNs they were active on, and whether they believe bots influence their or others' behavior and why. To motivate participants to submit responses that best represent their perceptions, we emphasized that the study does not aim to test them and reminded them to answer honestly.

The next part aimed to gauge how accurately OSN users can differentiate between bots and non-bots. We presented participants with 20 carefully chosen Twitter accounts and asked them to classify each as a bot or a non-bot. This part of the study consisted of two stages. In each stage, we asked participants to classify ten accounts, split equally between bots and non-bots. Between the two stages, we showed a simple, concise definition of bots and their goals (inspired by Grimme et al. [27] and Ferrara et al. [23]), as a means to inform participants who are unfamiliar with bots, and measure whether informing participants about bots improves classification accuracy.

To conclude the study, we asked participants to share their level of satisfaction with how the OSNs they have accounts on deal with bots (five-point Likert scale), and asked them how OSNs should best use bot-detection technology to moderate bots. Finally, we collected demographic information.

We decided to ask participants to classify Twitter accounts rather than accounts from other OSNs since Twitter bots are well-studied, and there are numerous public datasets we could select accounts from. We selected accounts from the 14 datasets used by Yang et al. [58]. We randomly ordered the accounts available in all datasets and manually inspected individual accounts one by one to select active ones pertaining to different types identified in prior work [19, 27, 35, 51]. Specifically, we selected a diverse set of bot and non-bot account-types (benign bot, political bot, …) differing in various features, such as their popularity or the type of content they post, to measure how account characteristics affect participants' classification. Overall, we selected 20 accounts: two groups of ten accounts each, split equally between bots and non-bots of diverse types. Tab. 1 presents the accounts and their characteristics. We presented participants with the accounts of one group in the first stage, followed by the accounts of the other group in the second stage. To mitigate ordering effects, we alternated between presenting group *A* first followed by group *B* and vice versa between stages, and randomized the account order within each stage. Each classification question contained a screenshot of the account's profile page (containing picture, description, number of followers, …), followed by as many of the account's most recent tweets as would fit on a 1,920×1,080 screen (a minimum of one tweet, and a median of 2), see Fig. 1.

Figure 1: A screenshot of a bot shown to participants in the study.

Cls	*ID: Definition*	*Group A*	*Group B*
Non bots	`Organization`: Organization or business	UHCF_CT	LGRforCollege
	`Parody`: Self-identifies as parody	notzuckerberg	Queen_UK
	`Popular`: >1,000 followers, unverified	aaroni	GynoStar
	`Unpopular`: <1,000 followers, unverified	kb_kinsey	bablackwood
	`Verified`: Of public interest, verified	DineshDSouza	TheBrianneDavis
Social bots	`Benign`: Entertains or serves users	EndlessJeopardy	TwoHeadlines
	`DateScam`: Publishes dating scams	brown_gabathaa	drrajashekar170
	`Fake`: Inflates number of followers	aurelias_alaura	navash73
	`Political`: Backs pol. agenda, attacks others	HansonMAGOP	shinziNET
	`TechScam`: Publishes technical scams	littlebilly_	lcj18344

Table 1: The accounts we asked participants to classify as bots and non-bots. The two account groups are listed on the right. The leftmost columns list the account classes, and type identifiers and definitions. Following best practices, we reported the malicious bots (last four rows) to Twitter.

Our study protocol (see the supplementary material) was reviewed and approved by our organization's ethics department.

3.2 Analysis

We qualitatively analyzed participants' definitions of bots and their goals—specifically, via inductive coding [45]—to understand participants' mental models. Over two sessions, two coders analyzed the responses of a subset of participants to agree upon a codebook. The coders later coded the responses of a subset of additional participants independently to assess the inter-coder agreement. Upon reaching substantial agreement, the coders split the coding effort of the remaining responses equally among themselves.

Subsequently, we built logistic regression models to estimate how various factors (mental models, account types, familiarity with Twitter, etc.) impacted participants' likelihood (i.e., odds ratio) of classifying accounts correctly. More precisely, because the responses to certain account-classification questions were *not* independent (each participant classified 20 accounts), we built mixed-effects models [47]—a family of models that can handle correlations between data points. For each explanatory variable, mixed-effects logistic regression models estimate how much an increment of one in explanatory variables affects the expected odds of correct account classification compared to a pre-specified baseline.

3.3 Participants

To determine the number of participants needed, we ran a pilot with a convenience sample (N=29). A subsequent power analysis (α=0.05, power=0.80) indicated that ≥288 participants were needed to confirm various factors' impact on classification accuracy.

We administered the study via Qualtrics and recruited participants via Prolific—an online crowdsourcing platform. To avoid self-selection bias, we did not advertise the study as one about bots. Instead, to remain inclusive, we advertised it as one aiming to learn about experiences on different OSNs and explore usage habits. Compared to other platforms (e.g., Amazon's Mechanical Turk), Prolific's participants are known to produce data of higher quality [40]. To further enhance data quality, we used Prolific's pre-screening options to limit participation to workers fluent in

English (the study's language) whose submission approval-rates are higher than 95%, thus obviating the need for attention questions [41]. Furthermore, we limited participation to 18-year-old or older workers who own an account on at least one of the seven OSNs our study asked about. A total of 298 participants matching these criteria completed the study. We excluded one participant due to submitting off-topic responses, thus remaining with 297 participants for the analysis. Participants took an average of 22 minutes to complete the study and were compensated $2.60 for their time.

Our participants were most active on Facebook, Instagram, and YouTube—≥86.19% reported owning accounts on at least one of these OSNs. A substantial portion reported owning accounts on Twitter (62.96%), LinkedIn (37.71%), or Reddit (53.20%). TikTok was least popular among participants (29.29% owned accounts). Our participants were skewed toward younger and male populations. Ages ranged between 18 and 61, with a median of 23. Males constituted 65.32% of the total participants (34.01% reported female, and the remainder chose other). The participants were educated (29.63% college students, 44.11% owned an associate's degree or higher) and mostly self-reported being technology savvy (76.43%). They lived in different parts of the globe, but mainly Europe (82.83%), North America (11.11%), and South America (4.71%).

3.4 Limitations

Some limitations should be considered when interpreting our results. The account-classification task in our study is Twitter-centric. Thus, results may not precisely reflect how well users can tell apart bots and non-bots on other OSNs. Nonetheless, we attempted to interpret our results generally and believe that our recommendations (§5) apply to all major OSNs. We also stress that our findings about user (mis)perceptions are OSN-agnostic. We further note that some participants (37.04%) reported not owning Twitter accounts, which could have disadvantaged them in the classification task. We controlled for lacking Twitter accounts in the analysis (§4.2) and found it had no significant impact on classification accuracy.

Unlike our participants, OSN users may not consciously seek to differentiate between bots and non-bots. Thus, our estimate for how well users detect bots in practice may be imprecise. Still, we believe our estimate is useful, as it likely sets an upper bound on how well OSN users perform in reality. Said differently, we expect that OSN users who do not consciously attempt to detect bots are likely to detect them less accurately than our participants.

We asked participants to classify accounts based on screenshots of profile pages, thus not faithfully imitating real browsing settings. Yet, we believe that our study design captures many occasions in which users consume content on OSNs via cursory examination without gathering additional information about accounts [25].

Lastly, as is common in online self-reported surveys, ours is not free of biases [28, 31]. Primarily, participant responses (e.g., on OSN use) may be affected by recall or desirability biases, or by misinterpreting questions. Additionally, our participants do not fully represent the general OSN user base—they were skewed toward the young, educated, and male population. We mitigated biases via careful study design and piloting, and reminding respondents to answer thoroughly and honestly. Furthermore, to validate that our findings hold despite potential biases, we repeated our study with different participants and reached consistent conclusions (§4.4).

4 RESULTS

Now we turn to the results. We start with participant perceptions of bots and how they influence users. We then report participants' account-classification accuracy and discuss factors impacting it. Next, we turn to participant estimates of bot prevalence, and discuss how these relate to their satisfaction with how OSNs govern bots. Finally, we report validation results with additional participants.

4.1 Mental Models

What are social bots? Most participants (79.80%) self-reported being familiar with the notion of bots. We asked those participants to define bots and list their goals, while we encouraged the others to make best guesses. The responses were instructive and reflected varied mental models about bots; we leveraged inductive coding to characterize them. Initially, two coders developed a codebook over two coding sessions. After analyzing 40 responses, they concluded the initial open coding process, as new themes stopped emerging. The coders then independently coded additional 40 responses to calculate the agreement among them, reaching substantial agreement (Cohen κ=0.76) [37]. Subsequently, the coders equally split the remaining responses between them for coding. Tab. 2 presents the final codebook along with example phrases, representative quotes, and the percentages of participants for which codes apply.

Participants mentioned several properties to describe bots. In most cases, the definitions captured actual bot properties. A large portion of the participants (49.83%) indicated automation as part of definitions, using terms such as artificial intelligence and algorithm to describe how bots operate. A substantial chunk of participants (20.20%) described bots as "fake" or ungenuine accounts without specifying how they are controlled, often referring to so-called fake followers that users can buy to promote account popularity [48]. Some even revealed that they or someone they know have once purchased fake followers. Various participants (19.19%) described bots as interactive accounts that communicate or interact with other accounts, while some (16.50%) noted that bots often pretend to be humans. A few participants (3.70%) suggested that bots use specific technologies to produce original content, and a smaller set (1.01%) described bots as anonymous users. Other participants provided vague or inaccurate definitions. A substantial amount of participants defined bots in abstract terms (17.51%), describing them as "something" that performs specific tasks, while certain participants provided cyclical definitions (12.79%). Some participants (6.40%) confused bots with other technologies, primarily cookies and other tracking mechanisms, and a few participants (2.02%) thought bots were sock puppets—users that masquerade as others for deception [8]. Lastly, a few participants stated they do not know what bots are (2.36%) or provided unintelligible responses (1.68%).

The participants articulated varied goals, reflecting the real-life diversity in bot usage. Multiple participants (42.09%) suggested that bots sought to artificially boost or inflate account or post popularity. Many (31.31%) stated that bots aim to influence users or sway public opinion (e.g., via disinformation). Some (11.11%) thought of bots as benign actors providing features to help users—ranging from entertainment to automated question answering—with a subset specifically citing customer support (6.40%) and legitimate advertising (9.09%). Participants also identified that bots can be used to

Cat.	*Code*	*%*	*Example(s)*	*Quote*
Definition	Automated	49.83%	AI; agent; program	P145: "A form of AI that contributes to a discussion."
	Fake	20.20%	"Fake"; ungenuine	P137: "It's a fake account that can create content or interact with other users."
	Interactive	19.19%	Communicate; interact	P170: "An automatic account that interacts with other accounts, for example makes comments."
	Abstract	17.51%	"Something"	P139: "A bot is something that posts or interacts with other people online."
	Impersonate	16.50%	Pretend to be human	P157: "[…] pretends to be a user, mimicking human behavior on Internet."
	Cyclical	12.79%	Social bot; bot	P198: "It is an account that is managed by a bot…"
	Confuse	6.40%	Cookie; trackers	P37: "Are they like clever cookies that influence the content you see?"
	Create content	3.70%	Create original content	P284: "It is designed to create posts. Mainly used as automated replies."
	Don't Know	2.36%	-	P149: "I don't know what is social bot …"
	Sock puppet	2.02%	Masquerade as others	P121: "[…] humans […] running many accounts […] with the purpose of persuading other genuine users."
	Unintelligible	1.68%	-	P281: "Social bot is not completed project nowadays."
	Anonymous	1.01%	Anonymous user	P271: "…an anonymous 'user,' which […] boosts up the popularity of certain accounts, persons, politicians."
Goal	Boost	42.09%	Inflate popularity	P195: "[…] bots could be used to e.g. artificially boost popularity of celebrities."
	Influence	31.31%	Sway public opinion	P174: "To influence your opinions."
	Benign	11.11%	Help users	P26: "[…] communicative agent that can be called in the conversation to give you any particular info."
	Confuse	10.44%	Personalize; recommend	P167: "Making things happen automatically by using certain algorithms, like chosing who will get which ads."
	Scam	9.76%	Spam; commit crime	P156: "Some try to trick you in clicking on malicious links."
	Advertise	9.09%	Promote products	P264: "[…] to interact with people with some content for e.g with some advertisment about some product."
	Earn money	8.08%	Make money	P215: "[Bots' goals are] Political influence or monetary gain."
	Support	6.40%	Customer support	P12: "Make the customer service work easier for companies…"
	Don't Know	4.71%	-	P242: "Sincerely, I don't know the goals of such social bots."
	Unintelligible	4.04%	-	P281: "I believe social bot is a good friend of all people."

Table 2: Codes describing mental models of bots. The second column from the left reports the percentage of participants for which codes apply; the third column lists examples of common words and phrases used by participants; and the rightmost column provides representative participant quotes. Codes highlighted in gray describe participants' responses, not bots.

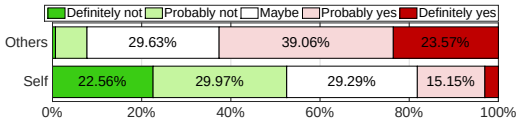

Figure 2: Participants believed that their own behavior was not influenced by bots, but that others' was.

scam OSN users and spread spam (9.76%), and are often operated to earn their operators monetary gains (8.08%). A substantial group (9.76%) confused bots' goals with those of other technologies, such as recommending content to users and personalizing the user experience. Lastly, a few participants were unable to articulate the goals of bots (4.71%) or submitted unintelligible responses (4.04%).

Who is influenced by bots? We asked participants whether they thought bots influenced their or others' behavior on OSNs. Fig. 2 presents the results. Most participants (52.53%) thought that bots were unlikely to influence them. To explain why, participants mentioned that they do not interact with bots, do not believe everything they read, are challenging to mislead, and can spot bots accurately. This is consistent with Geeng et al.'s finding that most people do not believe they are susceptible to propaganda and disinformation [25]. Somewhat paradoxically, however, participants misclassified a significant portion of bots as non-bots (§4.2), indicating that bots are likely to unconsciously deceive them in reality.

By contrast, most participants (62.63%) believed that others were likely influenced by bots. Participants explained their belief by noting that others likely do not know what bots are, cannot distinguish bots from non-bots, and are likely to be misled into thinking that certain content is more popular than it truly is.

4.2 Users' Ability to Detect Bots

Accuracy. We examined how well participants could distinguish between bots and non-bots. Overall, participants classified 68.30% (±12.68%) of the accounts correctly. Namely, participants correctly classified about 14 of the 20 accounts on average, with most participants (76.43%) making between 11 and 17 correct classifications. At the same time, none could classify all 20 accounts correctly, and many participants (10.10%) misclassified ≥50% of accounts.

When examining misclassifications, we found that participants were more likely to misclassify bots as non-bots than the other way around. More specifically, participants misclassified an average of 23.27% of non-bots as bots, and 40.13% of bots as non-bots. This is worrisome, because by tricking participants into believing they are non-bots, malicious bot owners are more likely to accomplish their goals (e.g., carry out a successful scam).

Factors. We built a mixed-effects regression model to assess how various factors impacted the participants' likelihood of classifying accounts correctly. The dependent variable in the model was an indicator for whether a participant answered a certain classification question correctly (1 for a correct answer, 0 otherwise). We derived the explanatory (a.k.a. independent) variables from the survey responses and our qualitative analysis. We used a random intercept as a random effect in the model to account for differences between participant baseline likelihoods to classify correctly.

Initially, we incorporated a rich set of independent variables in our model, including ones describing *1)* the OSNs participants were heavily active on (e.g., spending at least an hour a day); *2)* whether participants indicated being familiar with bots; *3)* the study completion time; *4)* the account type being classified (dating scam bot, verified non-bot, …); *4)* participant demographics; *5)* mental models of bots and their goals; *6)* whether participants saw the bot definition we provided (i.e., whether they were in the first or second stage of the classification task); *7)* interactions between whether participants saw the definition and other variables (account type, familiarity with bots, …). The initial model and variables are available in the supplementary meterial. Next, we followed a standard

Category	*Variable*	*Odds*	*Conf. int.*		*p-val.*
Intercept	Mean	3.10	[2.44,	3.94]	**<0.01**
	Std	0.30	-	-	-
Demographics	AgeScaled	0.63	[0.39,	0.99]	**0.04**
	HighDegree	1.31	[1.14,	1.52]	**<0.01**
Habits	HeavyReddit	1.38	[1.17,	1.62]	**<0.01**
Perceptions	DefAbstract	0.76	[0.63,	0.94]	**0.01**
	DefAutomated	0.84	[0.71,	0.98]	**0.03**
	DefImpersonate	1.30	[1.06,	1.57]	**0.01**
	GoalDontKnow	0.67	[0.49,	0.91]	**0.01**
	GoalScam	1.43	[1.13,	1.82]	**<0.01**
	GoalUnintelligible	0.65	[0.47,	0.91]	**0.01**
Account Type	BotBenign	0.66	[0.51,	0.87]	**<0.01**
	BotDateScam	0.84	[0.64,	1.08]	0.17
	BotFake	0.19	[0.15,	0.24]	**<0.01**
	BotPolitical	0.47	[0.34,	0.64]	**<0.01**
	BotTechScam	0.58	[0.45,	0.75]	**<0.01**
	NonBotParody	0.58	[0.45,	0.75]	**<0.01**
	NonBotPopular	2.44	[1.79,	3.29]	**<0.01**
	NonBotUnpopular	1.51	[1.14,	1.99]	**<0.01**
	NonBotVerified	1.35	[1.03,	1.77]	**0.03**
Stage	SawDef	0.84	[0.74,	0.97]	**0.01**
	SawDef:BotPolitical	1.52	[1.06,	2.18]	**0.02**

Table 3: Logistic regression model's parameter estimates after model selection. For each variable we provide the estimated mean odds, the 95% confidence intervals, and the p-value for the null hypothesis that the mean odds are equal to 1. Except for `BotDateScam`, all mean odds are estimated to be different than 1 with statistical significance (p-value<0.05).

backward model-selection process to simplify the model while maintaining a good fit [47]. Specifically, we gradually removed the variables with the highest p-values, one at a time, until reaching a point at which the Bayesian Information Criterion (BIC) decreased by <5. Eventually, the process yielded the model with the variables listed in Tab. 3 alongside their estimates (R^2=15.81%, BIC=7010). The model explains 15.81% of the variance in the data, indicating that there are missing factors that can further explain classification accuracy. Nonetheless, in a 5-fold cross-validation process, we found that a predictive model using only the fixed effects can predict when participants classified accounts correctly or not with high accuracy (70.77% mean accuracy and 68.16% Receiver-Operating Characteristic area under curve).

The baseline in the model is an 18-years-old user with a high-school or lower education level, who does not use Reddit heavily, and who is classifying an organizational non-bot account before seeing the definition. From the intercept, we can see that the mean odds that the baseline user will classify the account correctly is 3.1. Said differently, the mean probability of correct classification is $\frac{3.1}{3.1+1} \approx 0.76$. Compared to the baseline, a user of the maximal age we encountered in our study (61) is estimated to have ×0.62 lower mean odds of correct classification (every year of age decreases the odds by about ×0.99), while a user with an associate's degree or higher is likely to have ×1.31 higher mean odds of correct classification. Other demographic properties (e.g., gender) and the study-completion time had no statistically significant correlation with classification's correctness (i.e., these variables were dropped by model selection).

The types of OSNs that participants reported using heavily had little impact on classification accuracy—except for the variable indicating heavy Reddit usage, none of the others survived model selection. Interestingly, despite asking participants to classify Twitter accounts, heavy Twitter users did not perform better than others. Heavy Reddit users, however, had ×1.38 higher mean odds of correct classification than the baseline. We hypothesize that this is due to Reddit users being particularly young and educated [42] and their constant exposure to benign bots [35].

The model also tells us that user perceptions of bots impact their classification accuracy. Notably, participants who provided abstract definitions of bots and ones who mentioned not being aware of bots' goals had ×0.67–×0.76 lower estimated mean odds for correct classification than the baseline. In contrast, participants who mentioned that bots may attempt to impersonate humans and mentioned scams as a potential goal of bots had ×1.30–×1.43 higher estimated mean odds than the baseline. This indicates that helping users form more accurate mental models of bots may significantly improve their ability to distinguish between bots and non-bots.

We can further learn from the model that the types of accounts being classified impacted participants' accuracy. Except for bots used for dating scams, participants' estimated mean odds of classifying bots correctly was ×0.19–×0.66 lower than the baseline, with fake followers being particularly difficult to classify correctly. By contrast, non-bots were generally easier for participants to classify correctly—participants exhibited ×1.35–×2.44 higher estimated mean odds to classify unpopular, popular, and verified non-bots correctly compared to the baseline. The exceptions are parody non-bots with significantly lower estimated mean odds (×0.58) of correct classification compared to the baseline. These findings highlight that aiding users in identifying bots via specialized interventions targeting specific types of accounts could be a promising avenue.

Finally, we can learn from Tab. 3 that seeing the definition before the second stage of account classification did not improve participants' performance: while the estimated mean odds of correctly classifying political bots increased by ×1.28 after participants saw the definition, the odds decreased by ×0.84 in all other cases. Potential explanations are that the specific definition we provided was not sufficiently accessible to the participants, repeated exposures to definitions might be needed to improve participant accuracy, or that participants paid slightly less attention to questions later in the study. Thus, further research is needed to find out how to develop an accessible definition of bots to help users identify such accounts in practice and determine the definition's effectiveness.

Classification heuristics. Participants relied on various rules of thumb to detect bots. Certain participants based their classifications on simple rough statistics from account *metadata*, such as the number of tweets and followers, the ratio between the number of accounts followed and the number of followers, the amount of retweeting of others' posts, and the number of hashtags in tweets. Other participants mainly relied on the *contents* of tweets and profile pages, including whether they contain suspicious links, whether content included promotions and marketing material, the existence of grammatical and spelling mistakes in posts, and repetitiveness and originality of posts (e.g., whether they repeated titles of linked articles). Lastly, other participants relied on *appearance and feel* and stated that bot tweets seemed "unnatural" and "strange," their photos appeared seductive or unrealistic (e.g., in dating scams), and that their posts seemed "sketchy." The rough statistics used by participants were similar to those used by ML algorithms for

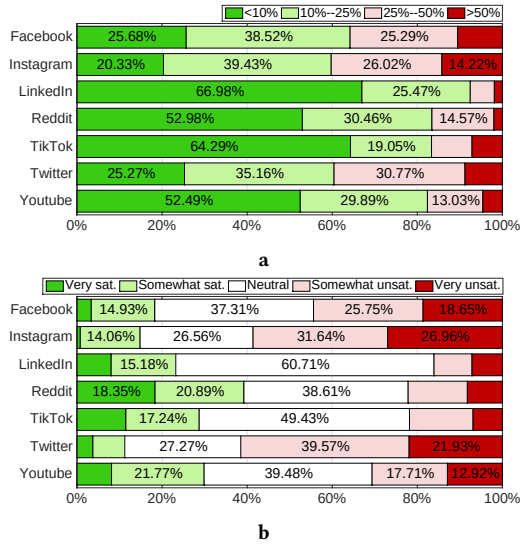

Figure 3: (a) Participant perceptions of bot prevalence (percentage of accounts) on OSNs. (b) Participant satisfaction with bot governance.

bot detection. Nonetheless, participant heuristics were generally less exhaustive, precise, and systematic than ML algorithms, thus helping explain participants' relatively high misclassification rates.

4.3 Perceptions of Bot Prevalence

When asked to estimate the prevalence of bots on the OSNs they own accounts on, participants believed they were most prevalent on Facebook, Instagram, and Twitter (Fig. 3a). Many participants (≥35.80%) expected that ≥25% of the accounts on those OSNs are bots. While participants likely overestimated the actual prevalence of bots—Varol et al. estimated that 9%–15% of active accounts on Twitter were bots [54]—their responses reflect that bots on certain OSNs are likely responsible for a substantial amount of the interactions with humans.

Participants expressed higher satisfaction levels with how OSNs governed bots when they perceived bots were less prevalent on the OSNs, and vice versa (Fig. 3b). Specifically, participants were less satisfied with how Facebook, Instagram, and Twitter dealt with bots than with the other OSNs. Unsatisfied participants attributed their lack of satisfaction to the lack of OSNs' efforts to moderate bots (e.g., by removing inappropriate adult content or propaganda), which they sometimes interpreted as lack of care for users; being frequently bothered by automated comments, direct messages, and connection requests; and preference to interact with humans. On the other hand, neutral participants expressed that it was unavoidable to have some bots on the platform; detecting bots should not entirely be OSNs' job, as users should develop a healthy skepticism toward the content they consume; and they do not encounter bots often. Finally, satisfied participants attributed their satisfaction to OSN efforts to moderate bots which they felt were tangible (e.g., OSNs swiftly removing content flagged as spam) and clearly publicized; the utility they found in (benign) bots; and to not being exposed to scams.

When asked about preferred courses of action OSNs should follow to moderate bots, participants were provided with several options for benign and malicious bots: suspending accounts, adding indicators identifying bots, or taking no action whatsoever. While systems and tools can conceivably perform a broader set of actions (e.g., move bot-produced content to a designated OSN section), we hoped that responses would shed light on whether there is a strong preference toward a specific course of action or whether different users might have different preferences for how OSNs should handle bots. The responses were heterogeneous, and participant preferences for malicious and benign bots were mostly different. Most participants (61.62%) stated that they would like detection tools to suspend malicious bots and allow benign bots while showing clear indicators identifying them as such. A large portion (21.21%) expressed that they would prefer OSNs to suspend all bots, regardless of whether they are malicious or benign, and another substantial portion (11.78%) expressed that they would prefer OSNs to allow all bots while showing indicators. While systems and tools can conceivably perform a broader set of actions (e.g., move bot-produced content to a designated OSN section), the results highlight the necessity of considering a additional bots-moderation options and accommodating the varied needs of OSN users (§5.2).

4.4 Additional Validation

Because our main set of participants was skewed, we recruited another cohort of participants to validate our findings' generalizability. Specifically, we recruited 139 volunteers from our organization—a global software company—to participate in the study. This validation sample was more balanced with respect to age (40 vs. 23 median age) and gender (55.40% vs. 65.32% male) than the participants recruited via Prolific. The validation participants resided mostly in North America (65.47%), Europe (20.14%), and Asia (13.67%). As expected, a large portion (37.41%) identified as computer engineers, but many identified as finance, legal, and administrative support professionals, among others. TikTok and Reddit (≤19.42% owned accounts) were the least popular OSNs among the participants. Most participants (59.71%) owned Twitter accounts, and ≥64.03% owned accounts on at least one of the remaining OSNs.

The validation participants exhibited similar mental models of bots as the Prolific participants and obtained equivalent account-classification accuracy (67.37%). Compared to the Prolific participants, the validation participants erred slightly more when classifying non-bots (31.29% misclassification), but did better on bots (33.96% misclassification). The relatively small sample size of the validation dataset precluded fitting a regression model on the dataset itself. Hence, to test whether the validation data yield parameter estimates similar to the Prolific data, we fit a mixed-effects logistic regression containing all parameters listed in Tab. 3, while adding an indicator variable for the datasets and an interaction term between the dataset and each of the parameters. The resulting model helped estimate whether the validation data yield different slopes or intercept compared to the original data. We found no statistically significant difference between the estimates of the intercepts and 15 of the 20 parameters across the two datasets, while three of the five remaining parameters had a similar correlation type (i.e., positive or negative) with the dependent variable for the two datasets. To

further increase our confidence in the findings, we tested how well can a model trained on the Prolific data using only fixed effects predict whether the validation participants accurately classified accounts. The model achieved 71.26% accuracy and 57.80% Receiver-Operating Characteristic area under curve—comparable to the performance achieved on the Prolific data. Finally, the validation participants' satisfaction with how the different OSNs dealt with bots was similar to that of the Prolific participants. Primarily, the participants were least satisfied with Facebook, Instagram, and Twitter. Thus, overall, we conclude that our findings on the Prolific data generalize well to our validation data.

5 DISCUSSION

Our study demonstrates that OSN users often misperceive bots and struggle in differentiating between bots and non-bots. Even worse, users are especially likely to mistake malicious bots as non-bots, thus rendering them susceptible to disinformation and attacks. Hence, there is a dire need for interventions to help users identify different OSN account-types. We now discuss potential education-based and technical interventions.

5.1 User Education

Our study shows that participants with abstract perceptions of bots and those unaware of bot goals were more likely than others to misclassify accounts. By contrast, participants familiar with the specificities of bots (e.g., usage to spread scams) detected bots with higher accuracy than others. These results indicate that user education and habituation to bots can significantly boost their ability to distinguish between bots and non-bots. We believe that education-based interventions can be useful, but they may be insufficient to prevent account misclassification, as even expert users who were familiar with bots misclassified accounts.

5.2 Systems and Tools

ML-based bot-detection algorithms differentiate between bots and non-bots with significantly higher accuracy than our participants. Using the same type of information that was available to our participants, algorithms often detect >95% of bots while misclassifying a small portion (<5%) of non-bots [32, 58]. In comparison, our participants misclassified 40.13% of the bots and 23.27% of the non-bots we asked about. The participants' lower accuracy can be attributed to using less comprehensive and systematic classification heuristics than algorithms. Additionally, bots can be abnormal in non-obvious ways; for instance, high follower growth-rates are useful features in bot detection [58]. Computing this and similar metrics is possible via OSN-provided data, but doing so manually for many accounts is infeasible. OSNs can help users detect bots by making such metrics visible, thus expanding the limited feature set they typically rely on. More generally, ML-based algorithms can empower users to detect bots accurately while avoiding disinformation and attacks.

Where to deploy detection? Detection algorithms can be deployed at various locations. One possibility is to make them available via OSN-independent websites, such as BotOMeter [57], that users would visit to classify certain accounts, or browser extensions, such as BotSight [39], that annotate accounts as bots and non-bots in situ, as users browse OSNs. This deployment model suffers from a scalability issue. Because users need to visit dedicated websites or install extensions actively, the user base of OSN-independent tools tends to be relatively small, thus leading to a limited impact. For example, out of the hundreds of millions of OSN users, the BotOMeter website had ~500 daily visits as of two years ago [57]. We also expect that the users who are likely to benefit the most from detection algorithms—those who are less familiar with bots or falsely believe that they can accurately detect bots—are unlikely to use OSN-independent tools.

Another, preferable, possibility is to deploy detection algorithms internally, behind the scenes, on OSN platforms. Unlike in the previous deployment model, all OSN users would benefit from the detection results. Indeed, certain OSNs like Reddit rely heavily on bot-detection systems to detect and suspend malicious bots [43]. We found that users of such OSNs are more satisfied with how bots are dealt with compared to other OSNs that are known to have a lax policy towards bots.

How to use detection results? Once they detect bots, systems and tools may react in various ways. For example, they might suspend the accounts [50] or annotate their posts as automatically produced or erroneous [7, 39]. The lack of consensus between our participants on the preferred course of action and the different preferences for malicious and benign bots both indicate that OSNs would better serve users by providing them with controls over how to treat different bot types. Currently available systems and tools, however, typically cannot distinguish between malicious and benign bots, and are programmed to take a single predetermined action users cannot adjust (e.g., adding indicators to all bots [39]). Additionally, bot-detection algorithms are typically evaluated on various accounts coarsely categorized as bots and non-bots, regardless of the specific type. These algorithms can potentially support users better if they help users detect specific account types that they are likely to misclassify (e.g., fake followers) with higher accuracy.

6 CONCLUSION

We studied 297 OSN users to improve our understanding of how they perceive social bots and measure their ability to detect them. While we found that participants' mental models often capture fundamental aspects of bots, some participants did not know what bots are, had an abstract understanding of bots, or confused them with other phenomena. Our study highlights participants' difficulty distinguishing between bots and non-bots: when asked to classify 20 accounts, all participants misclassified one or more accounts, with 10.10% of the participants misclassifying at least ten accounts. Mental models (e.g., abstract understanding of bots) and types of accounts being classified (e.g., fake followers) played central roles in participants' ability to classify accounts correctly. The participants showed a lack of satisfaction with how certain OSNs govern bots and expressed a desire for varied intervention types. Our findings help inform us about the benefits of educational and technical interventions and ways to improve them.

ACKNOWLEDGMENTS

This work was partially supported by a gift from KDDI Research, Inc.; and by Len Blavatnik and the Blavatnik Family foundation.

REFERENCES

[1] Norah Abokhodair, Daisy Yoo, and David W McDonald. 2015. Dissecting a social botnet: Growth, content and influence in Twitter. In *CSCW*.

[2] Abdulrahman Alarifi, Mansour Alsaleh, and AbdulMalik Al-Salman. 2016. Twitter turing test: Identifying social machines. *Information Sciences* 372 (2016).

[3] Seyed Ali Alhosseini, Raad Bin Tareaf, Pejman Najafi, and Christoph Meinel. 2019. Detect me if you can: Spam bot detection using inductive representation learning. In *WWW Companion*.

[4] Eiman Alothali, Nazar Zaki, Elfadil A Mohamed, and Hany Alashwal. 2018. Detecting social bots on Twitter: A literature review. In *IIT*.

[5] Darren Scott Appling and Erica J Briscoe. 2017. The perception of social bots by human and machine. In *IFC*.

[6] Adam Badawy, Kristina Lerman, and Emilio Ferrara. 2019. Who falls for online political manipulation?. In *WWW Companion*.

[7] BBC. 2020. Twitter tags Trump tweet with fact-checking warning. https://bbc.in/34Weicy. Last accessed on 06-19-2022.

[8] Zhan Bu, Zhengyou Xia, and Jiandong Wang. 2013. A sock-puppet detection algorithm on virtual spaces. *Knowledge-Based Systems* 37 (2013), 366–377.

[9] BuzzFeed. 2020. Whistleblower says Facebook ignored global political manipulation. https://bit.ly/33zB0GB. Last accessed on 06-19-2022.

[10] Ana Paula Chaves and Marco Aurelio Gerosa. 2021. How should my chatbot interact? A survey on human-chatbot interaction design. In *HCI*.

[11] Nikan Chavoshi, Hossein Hamooni, and Abdullah Mueen. 2016. DeBot: Twitter bot detection via warped correlation.. In *ICDM*.

[12] Maxime Clément and Matthieu J Guitton. 2015. Interacting with bots online: Users' reactions to actions of automated programs in Wikipedia. *Computers in Human Behavior* 50 (2015), 66–75.

[13] Stefano Cresci. 2019. Detecting malicious social bots: Story of a never-ending clash. In *MISDOOM*.

[14] Stefano Cresci. 2020. A decade of social bot detection. *Commun. ACM* 63, 10 (2020), 72–83.

[15] Stefano Cresci, Roberto Di Pietro, Marinella Petrocchi, Angelo Spognardi, and Maurizio Tesconi. 2015. Fame for sale: Efficient detection of fake Twitter followers. *Decision Support Systems* 80 (2015), 56–71.

[16] Stefano Cresci, Roberto Di Pietro, Marinella Petrocchi, Angelo Spognardi, and Maurizio Tesconi. 2017. The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race. In *WWW*.

[17] Clayton Allen Davis, Onur Varol, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. 2016. BotOrNot: A system to evaluate social bots. In *WWW Companion*. 273–274.

[18] Raquel Maria Dillon. 2022. A former employee accuses Twitter of big security lapses in a whistleblower complaint. https://n.pr/3DG9Hyh. Last accessed on 09-19-2022.

[19] Juan Echeverria, Emiliano De Cristofaro, Nicolas Kourtellis, Ilias Leontiadis, Gianluca Stringhini, and Shi Zhou. 2018. LOBO: Evaluation of generalization deficiencies in Twitter bot classifiers. In *ACSAC*.

[20] Manuel Egele, Gianluca Stringhini, Christopher Kruegel, and Giovanni Vigna. 2013. COMPA: Detecting compromised accounts on social networks.. In *NDSS*.

[21] Emilio Ferrara. 2020. Bots, elections, and social media: A brief overview. *Disinformation, Misinformation, and Fake News in Social Media* (2020), 95–114.

[22] Emilio Ferrara. 2020. What types of COVID-19 conspiracies are populated by Twitter bots? *First Monday* (2020).

[23] Emilio Ferrara, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini. 2016. The rise of social bots. *Commun. ACM* 59, 7 (2016), 96–104.

[24] David Mandell Freeman. 2017. Can you spot the fakes? On the limitations of user feedback in online social networks. In *WWW*.

[25] Christine Geeng, Savanna Yee, and Franziska Roesner. 2020. Fake news on Facebook and Twitter: Investigating how people (don't) investigate. In *CHI*.

[26] Katy Ilonka Gero, Zahra Ashktorab, Casey Dugan, Qian Pan, James Johnson, Werner Geyer, Maria Ruiz, Sarah Miller, David R Millen, Murray Campbell, et al. 2020. Mental models of ai agents in a cooperative game setting. In *CHI*.

[27] Christian Grimme, Mike Preuss, Lena Adam, and Heike Trautmann. 2017. Social bots: Human-like by means of human control? *Big data* 5, 4 (2017), 279–293.

[28] Andrew Guess, Kevin Munger, Jonathan Nagler, and Joshua Tucker. 2019. How accurate are survey responses on social media and politics? *Political Communication* 36, 2 (2019), 241–258.

[29] Taylor Hatmaker. 2018. This bot unrolls Twitter threads and turns them into readable blog posts. https://tcrn.ch/31pTsRY. Last accessed on 06-19-2022.

[30] Stefanie Haustein, Timothy D Bowman, Kim Holmberg, Andrew Tsou, Cassidy R Sugimoto, and Vincent Larivière. 2016. Tweets as impact indicators: Examining the implications of automated "bot" accounts on Twitter. *Journal of the Association for Information Science and Technology* 67, 1 (2016), 232–238.

[31] Ruogu Kang, Stephanie Brown, Laura Dabbish, and Sara Kiesler. 2014. Privacy attitudes of mechanical turk workers and the us public. In *SOUPS*.

[32] Sneha Kudugunta and Emilio Ferrara. 2018. Deep neural networks for bot detection. *Information Sciences* 467 (2018), 312–322.

[33] Srijan Kumar and Neil Shah. 2018. False information on web and social media: A survey. *arXiv preprint 1804.08559* (2018).

[34] Tetyana Lokot and Nicholas Diakopoulos. 2016. News Bots: Automating news and information dissemination on Twitter. *Digital Journalism* 4, 6 (2016), 682–699.

[35] Kiel Long, John Vines, Selina Sutton, Phillip Brooker, Tom Feltwell, Ben Kirman, Julie Barnett, and Shaun Lawson. 2017. "Could you define that in bot terms"? Requesting, creating and using bots on Reddit. In *CHI*.

[36] Luca Luceri, Ashok Deb, Adam Badawy, and Emilio Ferrara. 2019. Red bots do it better: Comparative analysis of social bot partisan behavior. In *WWW Companion*.

[37] Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica: Biochemia medica* 22, 3 (2012), 276–282.

[38] Amanda Minnich, Nikan Chavoshi, Danai Koutra, and Abdullah Mueen. 2017. BotWalk: Efficient adaptive exploration of Twitter bot networks. In *ASONAM*.

[39] NLOK. 2020. BotSight. https://bit.ly/2RHHv8c. Last accessed on 06-19-2022.

[40] Eyal Peer, Laura Brandimarte, Sonam Samat, and Alessandro Acquisti. 2017. Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology* 70 (2017), 153–163.

[41] Eyal Peer, Joachim Vosgerau, and Alessandro Acquisti. 2014. Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behavior research methods* 46, 4 (2014), 1023–1031.

[42] PRC. 2019. Who uses YouTube, WhatsApp and Reddit. *Pew Research Center* (2019).

[43] Reddit. 2020. Reddit Security Report. https://bit.ly/2SVJSBI. Last accessed on 06-19-2022.

[44] Elissa M Redmiles, Neha Chachra, and Brian Waismeyer. 2018. Examining the demand for spam: Who clicks?. In *CHI*.

[45] Johnny Saldaña. 2015. *The coding manual for qualitative researchers*.

[46] Giovanni C Santia, Munif Ishad Mujib, and Jake Ryland Williams. 2019. Detecting social bots on Facebook in an information veracity context. In *ICWSM*.

[47] Howard J Seltman. 2012. Experimental design and analysis.

[48] Indira Sen, Anupama Aggarwal, Shiven Mian, Siddharth Singh, Ponnurangam Kumaraguru, and Anwitaman Datta. 2018. Worth its weight in likes: Towards detecting fake likes on instagram. In *WebSci*.

[49] Mohammad Shafahi, Leon Kempers, and Hamideh Afsarmanesh. 2016. Phishing through social bots on Twitter. In *BigData*.

[50] Craig Timberg and Elizabeth Dwoskin. 2018. Twitter is sweeping out fake accounts like never before, putting user growth at risk. https://wapo.st/3lLlM8T. Last accessed on 06-19-2022.

[51] Milena Tsvetkova, Ruth García-Gavilanes, Luciano Floridi, and Taha Yasseri. 2017. Even good bots fight: The case of Wikipedia. *PloS one* 12, 2 (2017).

[52] Aman Tyagi, Matthew Babcock, Kathleen M Carley, and Douglas C Sicker. 2020. Polarizing tweets on climate change. In *SBP-BRiMS*.

[53] Joshua Uyheng and Kathleen M Carley. 2019. Characterizing bot networks on Twitter: An empirical analysis of contentious issues in the Asia-Pacific. In *SBP-BRiMS*.

[54] Onur Varol, Emilio Ferrara, Clayton A Davis, Filippo Menczer, and Alessandro Flammini. 2017. Online human-bot interactions: Detection, estimation, and characterization. In *ICWSM*.

[55] Kurt Wagner and Rani Molla. 2018. Facebook has disabled almost 1.3 billion fake accounts over the past six months. https://bit.ly/3knOsEA. Last accessed on 06-19-2022.

[56] Mairieli Wessel, Bruno Mendes De Souza, Igor Steinmacher, Igor S Wiese, Ivanilton Polato, Ana Paula Chaves, and Marco A Gerosa. 2018. The power of bots: Characterizing and understanding bots in OSS projects. In *HCI*.

[57] Kai-Cheng Yang, Onur Varol, Clayton Davis, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. 2019. Arming the public with artificial intelligence to counter social bots. *Human Behavior and Emerging Tech* (2019).

[58] Kai-Cheng Yang, Onur Varol, Pik-Mai Hui, and Filippo Menczer. 2020. Scalable and generalizable social bot detection through data selection. In *AAAI*.

[59] Jennifer Zamora. 2017. I'm sorry, Dave, I'm afraid I can't do that: Chatbot perception and expectations. In *Proc. HAI*.

[60] Xianchao Zhang, Shaoping Zhu, and Wenxin Liang. 2012. Detecting spam and promoting campaigns in the Twitter social network. In *ICDM*.