VoxPop: An Experimental Social Media Platform for Calibrated (Mis)information Discourse

Filipo Sharevski
DePaul University
Chicago, IL, USA
fsharevs@cdm.depaul.edu

Peter Jachim
DePaul University
Chicago, IL, USA
pjachim@depaul.edu

Emma Pieroni
DePaul University
Chicago, IL, USA
epieroni@depaul.edu

Nathaniel Jachim
University of Michigan
Dearborn, MI, USA
njachim@umich.edu

ABSTRACT

VoxPop, shortened for *Vox Populi*, is an experimental social media platform that neither has an absolute "truth-keeping" mission nor an uncontrolled "free-speaking" vision. Instead, it allows discourses that naturally include (mis)information to contextualize among users with the aid of UX design and data science affordances and frictions. VoxPop introduces calibration metrics, namely a *Faithfulness-To-Known-Facts (FTKF) score* associated with each post and a *Cumulative FTKF (C-FTKF) score* associated with each user, appealing to the self-regulated participation using sociocognitive signals. The goal of VoxPop is not to become an ideal platform—that is impossible; rather, to bring to attention an adaptive approach in dealing with (mis)information rooted in social calibration instead of imposing or avoiding altogether punitive moderation.

CCS CONCEPTS

• **Security and privacy → Social aspects of security and privacy**; **Usability in security and privacy**; • **Human-centered computing → Social networks**.

KEYWORDS

social media platform, misinformation, user incentive analysis, emergent moderation, inclusive usable security

ACM Reference Format:
Filipo Sharevski, Peter Jachim, Emma Pieroni, and Nathaniel Jachim. 2021. VoxPop: An Experimental Social Media Platform for Calibrated (Mis)information Discourse. In *New Security Paradigms Workshop (NSPW '21), October 25–28, 2021, Virtual Event, USA.* ACM, New York, NY, USA, 20 pages. https://doi.org/10.1145/3498891.3498893

NSPW '21, October 25–28, 2021, Virtual Event, USA

ACM ISBN 978-1-4503-8573-2/21/10.
https://doi.org/10.1145/3498891.3498893

1 INTRODUCTION

Human sociality is a fundamental concept that represents a dynamic relational matrix within which humans are constantly interacting in ways that are co-productive, continually plastic, and malleable [51]. Participating in this "matrix," then, was a natural preference for Internet users accustomed to the rudimentary repositories of threaded conversations like Usenet [89]. Initially referred to as "social network sites," the sociality achieved with multimedia affordances enabled these platforms to maintain pre-existing social networks and connect people based on shared interests, political views, or activities [13]. For a period, social media platforms were mainly seen as outlets for self-representation to wider audiences where the interaction was mostly centered around interpersonal discourse [24]. Then, malicious actors hijacked this discourse during the 2016 U.S. elections and the U.K. Brexit referendum to disseminate misinformation (or information *un*faithful to known facts) and inflammatory content [106]. Social media never recovered from these watershed events and, facilitated by populist accounts, became a go-to place for the dissemination of any type of misinformation—not just political "fake news" [84]. The already tense sociality was further exacerbated during the COVID-19 pandemic in which people turned to social media in the absence of definitive authoritative information about containing the virus and mass immunization [40].

In the wake of 2016, social media companies, politicians, and many constituents alike, realized that misinformation could not only affect election outcomes, fuel civil dissent, impact public health, but could also be detrimental to the platform's profitability [48]. Something clearly had to be done to curb misinformation, so Facebook started applying warnings on posts by adding "disputed" tags on stories that were debunked by fact-checkers, as well as fact-check tags under potentially misleading stories [82]. Twitter did not begin similar soft moderation until 2020, when, in late March, after the onset of the COVID-19 pandemic, the platform began issuing labels on tweets deemed as spreading misinformation related to the coronavirus [74]. Originally, only tweets that pertained to COVID-19 were flagged; however, following the 2020 U.S. presidential election in November, Twitter broadened the types of misleading, false, or disputed information to which it appended warning labels about the outcome of the election, claims of election fraud, or the safety of voting by mail (Twitter, together with Facebook, resorted to hard moderation of permanently banning accounts in the election

aftermath) [83]. Even Instagram followed suit and banned the accounts that spread COVID-19 vaccine misinformation with a similar approach to their big brother, Facebook [39].

Claiming infringement of the U.S. Constitution's "free speech" protections, some users were unhappy with the soft/hard moderation approach for "policing" the online discourse by mainstream players, and many people fled to alternative and low-profile social media platforms like Parler and Gab [41] (a similar "platform migration", although in the opposite direction, was noticed earlier from 4chan and Reddit toward Twitter [104]). The alternative social media platforms jumped at the opportunity to brand themselves as true "free speech" enablers, providing "public squares" for sharing any opinions. As it turned out, the free speech was directed toward discourse awash with misinformation unfettered by any counter-argumentation, as platforms amplified the voices of its pundits in the implicit role of "influencers" [61]. As Parler evolved and Gettr entered the picture, the resulting sociality, after a relatively short but tumultuous period, split into two camps with rather opposite sensitivity and receptivity to misinformation, rumors, and alternative narratives.

The division between "truth-keepers" and "free-speakers" created a rigid "us versus them" dichotomous choice of participation. Users could equally belong to both camps, but such participation seemed cumbersome with the current social media platform options. A user, for example, might accept the 2020 U.S. election outcome as legitimate and even support the winning candidate in the election (i.e. reject election misinformation), but be "unfaithful" to known COVID-19 facts due to a personal aversion to vaccinations [73]. Twitter realized that this might be a problem and backpedaled by introducing a striking system for COVID-19 misinformation in which accounts are disciplined based on the number of strikes the user has accrued for spreading COVID-19 misinformation (e.g., three strikes: 12-hour account lock, four strikes: 7-day account lock, and five or more strikes: permanent account suspension).

Another factor that contributed to this adaptive soft moderation is that social media platforms learned that misinformation labels could easily "backfire" or reinforce the user's belief in the misinformation [19]. Twitter has been experimenting with passing the buck to the community to employ soft moderation, instead of themselves, through their Birdwatch program [88]. This program allows for "fact-checkers" to write notes that provide context to the tweet, rate the quality of other participants' notes, and have these notes be visible directly on tweets with potential misinformation; this is available to all Twitter users (it is still in the experimental phase, facing the challenges of coordinated manipulation, bias, and harassment). Even Facebook realized that the commodification of "truth" is not a straightforward affair and intervention could also backfire, reversing their decision to suspend posts suggesting COVID-19 was man-made following President Joe Biden's directive to U.S. intelligence agencies to investigate competing theories on how the virus first emerged [68].

For mainstream social media companies, there is a series of conflicts that arises from the increased dissemination of misinformation online. First, to be considered *mainstream*, it implies the existence of an alternative *other* platform. By the mere existence of alternative platforms like Parler, Gab, and Gettr, Twitter and Facebook are able to appeal to the relative privation fallacy that mainstream sites are "not as bad" as the others and are thus outside the purview of regulation or criticism [72]. Second, social media companies also indiscriminately profit from user activity, regardless of the quality of narratives being disseminated [89]. This causes a dilemma, as engaging in moderation techniques could cause platform migration or diminish user interaction with the platform. Unless there is significant cause for platforms to substantially change by way of regulation or rugged capitalism, there is little hope for ensuring that constructive discourse is a public good. Therefore, we propose a novel and experimental platform, *VoxPop*, meant to offer a pragmatic response to the downfalls of both mainstream and alternative platforms, and which allows for the voice of the people, through careful calibration of VoxPop's sociality, to be heard.

2 VOXPOP PARADIGM

2.1 The Need for VoxPop

The diminishing nature of constructive discourse becomes even more apparent when the duopoly of social media platforms is seen from a *network society* perspective [16]. According to Castells, we exist in a society whose social structure is made up of networks such as the social platforms, and in such a society, the chief form of power is control or influence over communication. The struggle for power, therefore, creates a tension between the efforts of some platforms to impose their values and goals and the efforts of others to resist their domination. This tension reduced the online discourse to forms where (i) misinformation is excluded from critical engagement with and simple consumption of (mainstream); or (ii) information deemed faithful to known facts is included only to illuminate the consumptive value for misinformation (alternative). We refer to "information deemed faithful to known facts" as a content entropy counter to misinformation and avoid commodification of the word "truth" in the formulation of the VoxPop's paradigm, which we address in subsection 2.3.

VoxPop distinguishes itself from the mainstream-alternative social media landscape in that it allows, in adaptive form as described further in the paper, for misinformation to exist on par with any information deemed faithful to known facts. In that struggle for power in the network society, VoxPop instead allows for the users, not the platforms, to decide the degree of exclusion (or inclusion) of (mis)information for users to engage with and/or simply consume it, as well as the consumptive value of misinformation and information deemed faithful to known facts. While still a social networking platform, that allows for connecting with friends (like Facebook) and also public-square discourse (like Twitter and Parler), VoxPop re-envisions the way these concepts are implemented. It allows for both the friend-circle voted truth (Facebook model) and the pure up/downvoted truth (Twitter/Parler model) to exist on the platform while balancing out, with design and data science features, against pernicious faithfulness to known facts overall.

Unlike the mainstream social media platforms, VoxPop avoids assuming an authoritarian "truth-keeper" role because it does not resort to conventional soft or hard moderation but instead uses UX affordances and platform engagement to signal to users how much a piece of information is faithful to facts known at a given moment of time. The distinction here is that VoxPop's aims to harness the

democratic process of vetting facts that later might be reformulated, refuted, and appended without an absolute claims on truth and truthfulness. VoxPop is neither a "free speech" public square where anything goes because it incorporates so-called "design frictions" as points of difficulty occurring during interaction with the platform, prompting moments of reflection and more "mindful" interaction [22]. VoxPop, in essence, enables constructive *social calibration* of the platform's discourse. We define *social calibration* as a way of participation in a discourse or expressing one's opinion in consideration of social cues and not only on one's volition or in accordance with an agenda, political position, or bias. The social calibration is chosen as a replacement of the misinformation discernment - characteristic for the printed press through errata - corrections, and editorials as this seems impractical and slow (and could resemble the way that Twitter or Facebook approach their own official web pages/blogs).

Social calibration is a natural response to the Van Dijk's theory of discourse [90]. Both mainstream and alternative platforms fail to account for what Van Dijk considers *context*, or the relevant aspects in the social situation agreed upon by the participants themselves, and not imposed by a moderator (or lack of thereof). This imposition of context entails, and in fact we are witness of, that all users from particular social media platforms speak in the same way. In other words, current social media platforms give the illusion of, but do not actually deliver, the conditions for contexts to dynamically construct between users. The context of discourse become visible only indirectly when these constructs control interaction and discourse, which VoxPop allows, by shifting the construction power to the participants and away from the routine fact checkers, moderators, and crowd-sourced editors.

2.2 VoxPop's Appeal

VoxPop allows for users to be themselves—to vent, be sarcastic, or act as devil's advocate; it does not engage in content labeling, nor does it impose strikes about misinformation sharing. These rather explicit moderation techniques often backfire, (i.e., make people believe more, not less misinformation) [19, 57]. Even if they do not backfire and people confer with the suggestions from "fact-checkers," social media users sometimes might involve misinformation as part of their participation simply because is funny, and it will generate discussion among their friends (increase user's sociality) [102]. Unlike other types of unfaithful-to-known-facts content, e.g. spam or click-bait, misinformation has a particular value for one's self-determination on social media and group membership (i.e. spam and click-bait are irrelevant to users in that they seldom reflect desired values, preferences, and beliefs [15]).

VoxPop takes predisposition into consideration in the platform design to maintain a baseline acceptable decorum. The goal of VoxPop is not to be a neutral platform because that is impossible to achieve [35]. Instead it is in a state of perpetual balancing cycle of design/user testing/redesign of affordances, frictions, and reshaping in resonance with people's agreement on discourse contexts. VoxPop also takes into consideration the underlying psychology of misinformation [65]. People's faith in known facts might falter because they do not stop to reflect sufficiently on their prior knowledge (or have insufficient or inaccurate prior knowledge) [5]. Therefore, VoxPop does not *a priori* expect any level of reflexivity from its users, but creates conditions through its UX design and adaptive regulation of user engagement metrics where a user can move from intuitive ('System 1') to deliberative ('System 2') reasoning if the user, not the platform, wants to [63].

This "if" is a deliberate leap of faith by VoxPop in assuming that users *would* shift between reasoning when engaging in socially calibrated discourse. People come to social media to satiate their needs for autonomy, relatedness, competence, having a place, and self-identity [44]. None of these needs require use of 'System 2' reasoning in materializing one's self-determination on VoxPop, but we choose to offer such a possibility because socio-cognitive aspects also play an important role the dynamic construction of discourse contexts [90]. This is not to say that we are not aware of the cognitive manoeuvres aimed at selectively 'switching off' users' moral agency, allowing the engagement in misconduct that they user generally avoid on social media, known as a *moral disengagement* [59]. A user might not be aware of disengaging and causing harm, and when called out, might claim they did so in the name of their freedom of speech [8]. In response, VoxPop offers a selection of UX frictions that implicitly serve as *moral engagement* reminders without infringing on the users' freedom of speech while accounting for socially calibrated discourse (e.g. the"watermarking" and "moment of mindfulness" frictions shown in Section 4).

An important design aspect of VoxPop is that it also considers the natural phenomenon of homophily, e.g. the formation of echo chambers [34] as a "feature, not a bug". The mainstream platforms, through the design of their feeds and in the effort to maximize engagement, enable closed echo chambers of users that are hardly exposed to other perspectives [18]. The alternative platforms go even further to fence the echo chambers by even disabling the possibility for users to encounter other perspectives naturally (e.g., Parler does not allow search by keywords, but only by controlled subset of hashtags created and disseminated by platform's "influencers" [61]). VoxPop does not interfere if a natural echo chamber occurs on the platform, but shapes it to be transparent (e.g., users are exposed to other, contrasting perspectives from other "camps" by crafting the feed to include contrasting posts). Users can naturally encounter alternative perspectives or even misinformation if they want to, without the fear that the platform will tag them as the "other camp" (or show them adverts associated with the "others"). Exposure to opposing views on social media might increase political polarization in on certain topics [6], but VoxPop response of transparent echo chambers is precisely in eliminating the platform's imposition of context (in this case Twitter) as a factor of polarization.

2.3 A Brief Treatise on "Truth"

VoxPop paradigm's is novel in that it raises the question of how a social media platform could potentially do justice to the "truth" while considering the radical subjectivity of the "truth" as the watershed moment with the promulgation of information deemed unfaithful to known facts on social media circa 2015 [30]. Human sociality flourishes on trust; trust requires truthfulness; and truthfulness presupposes that there are (at least some) truths [96]. Following this simple genealogy, truthfulness has at least two virtues: (i) *accuracy* or the careful deliberation over the evidence (facts) for and

against a belief before assenting to it; and (ii) *sincerity* or genuinely expressing to others what one in fact believes.

Accurate and sincere reporting of truths is contextual to the discourse, in the view of VoxPop, and perhaps not strictly instrumental to the truthfulness of an argument. In a sincere and accurate discourse, participants are expected to arrive at truths about the way the world really is and that is what mainstream platforms brandish as their "truth-keeping" mission (of course, factoring for the conflicting needs to be able to amass enough content for monetization while not imposing too harsh moderation against inaccuracy and insincerity). But one can be inaccurate yet possibly sincere and leave the impression of deceiving the others. Take for example the mystery of the Havana Syndrome [9] or many invalided yet peer-reviewed academic papers [81]. Equally, one can be accurate to an extent but insincere; take for example the chronic immigration policy hypocrisy in the United States [21]. Social media posts akin to these examples, even during their (short) shelf life, served the purposes of trust and flourishing of human sociality so could passed as truths, even though many will count them as false.

In arriving at a consensus of truth, with the help of calculated platform moderation (or lack of thereof), the reporting of "truths" on social media became more than a participation of a discourse and any statement of faithfulness to known facts was assumed as a "trophy" of an argument: If one triumphs, another one must be defeated (if not humiliated), and there is a permanent record about it. The more "trophies" a user gets, the harder it gets for their "fans" (followers) to expect anything but victory in every next argumentative battle. Conversely, the more a user picks defeats on the "main stage", the more they look for other battlefields (Parler, Gab, 4chan, etc.) where they could become victorious. The reporting of truth on social media, at worst, is a reformulation of war and warfare on an interpersonal level. At best, it is a distorted materialization of the social identity theory and the in-/out-group dynamics [14]. As we noted, the self-determination pull towards social media is often bound with users' group memberships and because people view themselves positively rather than negatively, it follows that they will also seek to have a positive social identity from being members ("fans," "followers") of a truth reporting group. The evaluation of those group memberships, those social identities, is essentially comparative - so two camps of "truth-keepers" and "free speakers" social contexts emerge for evoking particular social identities in this battle of truth reporting.

Interweaving conflicting notions of "truth," aspects of social psychology and group dynamics, as well as the view of low-intensity interpersonal truth warfare is obviously too complex to answer how a social media platform could potentially do justice to the "truth" within this paper. But we start with a correspondence distinction between (i) "information deemed *faithful* to known facts" and (ii) "information deemed *unfaithful* to known facts." This formulation is selected to allow for richer dispositions to facts over time and to acknowledge that facts and dispositions are subject of change. This distinction does not avoid formation of groups on VoxPop and we expect that such a thing will happen; What does VoxPop hopes to avoid, instead, is a trench warfare on truth reporting. One could be inaccurate yet sincere or vice versa, and VoxPop novel paradigm is that it enables one to recover from such fluctuations without their self-determination suffering unrecoverable losses.

3 VOXPOP DESIGN: AFFORDANCES

3.1 Home Page

The proto-design of VoxPop's affordances, for brevity, are described only for browser interaction in the reminder of this section. A smartphone VoxPop application is critical for the platform's adoption, and we work in parallel on it. We decided for such a design approach as to critically revise and contextualize according to the VoxPop's paradigm the most prevalent psychological/economic mechanisms built in social media apps: (1) endowment effect, (2) social comparison and social reward, and (3) the "need for cognitive closure" [54]. As an alternative to these hallmarks of current social media platform design we hope to center our design around incentives to socially calibrate discourse around known facts, akin to the Wikipedia model of user-driven content moderation.

When VoxPop is opened in a browser, the user lands on a homepage as shown in Figure 11 or 12 in the Appendix. Users may first notice the color of the banner: either red or green depending on the user's current overall "Faithfulness-To-Known-Facts (FTKF) score". This choice is US-centric only to convey the idea of color contrast and considers these two colors only in that matter. Obviously, VoxPop allows for these colors to change (or perhaps only shades of one color) based on accessibility (e.g. colorblindness) or cultural context (e.g. red is considered a lucky color in China). Some social media platforms already do offer contrasting themes, for example 4chan, the notorious alt-platform, allows users to select several color themes (default is red) [61]. The contrast is important because it helps mitigate the risk of habituation where users that do not agree with the color scheme consider it irrelevant [45]. The hue utilized in the banner is also consistently used in the Calibration Dashboard segment, identified as "Metrics", which displays series of statistics based on the "Cumulative FTKF (C-FTKF) score" of the user logged in, as well as averages of FTKF scores for different demographics of VoxPop users (further explained in Section 5).

All elements throughout the home page of VoxPop utilize rounded edges, to give the platform a softer, more welcoming, and approachable feel than jagged or sharp edges [3]. On the right side of the home page view, is an element dedicated to "Other Voices" or voices on the platform that the user has decided to observe or listen to. On the left side will be two highlighted profiles: "In-the-green-zone" and "In-the-red-zone" chosen based on their week's relative change in their C-FTKF scores. These two highlighted accounts vary for each individual user based on the user's closest group of associates on the platform and which of those "friends" experienced the highest increase in the C-FTKF score from the previous week and which one experienced the lowest overall drop in the score from the previous week. Since the scores are relative to the previous week, it is unlikely to have repeated accounts in either position from week to week—if the same user occupies either "zone" for multiple weeks in a row, VoxPop may shuffle and highlight profiles in the 99^{th}-96^{th} percentile of users with the largest increases/decreases in weekly cumulative faithfulness to known facts.

These two elements are deliberately selected as placeholders for highlighting other users of interest as a measure of unique cross-pollination to VoxPop [92]. We understand and acknowledge it is

possible that a user may feel "bullied" by this design choice of VoxPop in that VoxPop exposes a user to unknown social groups that could potentially be perceived as unwanted, toxic, or in some cases, irrelevant. We believe, and will strive in the later design versions of VoxPop to make such cases rare exception because: (1) we incorporate "In-the-green-zone" and "In-the-red-zone"' elements in good faith, (2) we are devoted to resolve any conflict if reported by users; and (3) work with users to algorithmically avoid such perception or notions of bullying and targeted exposure as the discourse evolves.

VoxPop understands, attempts, and will attempt to avoid any loss of self-determination or harm to users that could come in two forms, internal (feelings, thoughts, doubts, regrets) and external (harassment by other users as a consequence of having been singled out in one of these elements). The naming is subject of change not just during the design but also during later versions of the platform. For example, both placeholders could be merged and renamed to a "Random-User-At-Week" in order to give an opportunity for exposure to users not recommended by the algorithmic tendencies for homophily of the mainstream social media platforms nor by the go-to influencers from the alt-platforms. Color scheme wise, the choice of red/green is again taken as a starting placeholder to allow for contrast based on the user's current overall FTKF score.

3.2 Color Scheme

Color informs the way we understand our surroundings, but the specific emotions associated with any given color are dependent on the environment or context that the color is presented in [27]. Our goal was to create two opposing color schemes that shift, dependent on the user's FTKF score, a concept elaborated on in Section 5. Multiple mainstream media platforms use blue as their primary color, including both Twitter and Facebook, which has been shown to be associated with positive connotations (e.g., openness, peace, calm, truth) whereas red, Parler's dominant color, has been shown to be associated with negative connotations (e.g., aggression, danger) [27, 61]. However, blue elements may be especially hard for the eye to pick up on a page [29], and since we wanted to utilize the two dominant colors in textual features, we decided to avoid it. Additionally, we wanted to steer away from the potential U.S. political associations of a red/blue color scheme, so while maintaining the red, we chose green as the "faithful" side of the color scheme which is thought to have similar connotations as blue when perceived [27] (the idea is to convey a *contrast* as a alternative "signal" to users about changes in their "voice").

For VoxPop we chose a soft grey hue as the background color that is less stark than pure white, in the hopes of avoiding eye strain in users [3, 95]. In addition to the grey, we chose a medium green hue that has sufficient contrast on the grey background, in order to ensure readability, since we intended to highlight profile names and the "Sourcing Friction" in the same green hue [29]. The bright red hue chosen was intended to highlight negative information and as such, is high-contrast relative to the grey background, and high in saturation relative to the green hue [29]. Red/green in and of itself is not friendly to color-blind individuals, so in future adaptations of VoxPop, we intend to consult with accessibility experts in order to include options to switch to a color-blind mode that will still make use of two dominant hues in the color scheme. In future design iterations, we may also choose to include a dark mode option that would utilize the same red and green features but placed over a dark background with contrasting text.

3.3 Calibration Dashboard

On the main page of VoxPop, the left side includes a Calibration Dashboard which includes metrics about the user and the user's associates on the platform, as shown in Figure 1 and expounded on in Section 5. The dashboard tracks the FTKF score of the user's last post and displays statistics based on the user's C-FTKF score in a palatable fashion through several different contexts. First a graph displays the C-FTKF scores of the user's closest "friends" on the platform—those users that are most interacted with. We chose this to provide the opportunity for the user to see for themselves, or reflect on, how they fare amongst their friends. Next, the dashboard displays the user's C-FTKF score relative to the friend-group to the second degree (i.e., friends of friends). This choice provides the user with a higher overview on their C-FTKF score relative to a more diverse group of VoxPop users. Another graph illustrates the regional C-FTKF scores based on geographical location of the user and relative to the user's own score, which is shown as forth. We added this level of comparison for the users to be able to track and compare trends in the voice of their "local people" over a week, but it could be over any period. The fifth and last metric before the user's C-FTKF is shown is the overall C-FTKF score of all active VoxPop users at the moment.

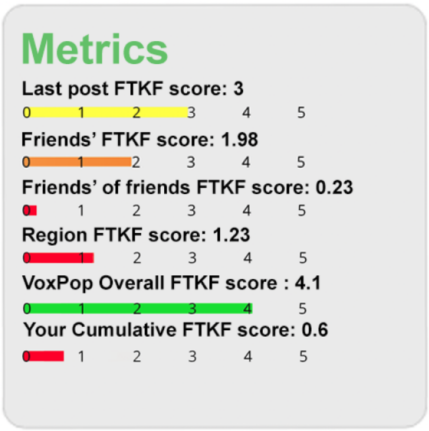

Figure 1: The calibration dashboard on VoxPop shows an individual user's metrics relative to other users in their extended social circle, averaged per circle.

Ultimately, the idea behind displaying C-FTKF scores is to play on the sense of social proof in users, how their group membership generally fairs, as well as how the particular context of the VoxPop discourse looks on different strata (a resembling idea, though without the scores, is presented to Twitter users around big events, e.g., the Oscars in which "friends" are "there" in the discourse) [36]. However, we are careful with the calibration dashboard and acknowledge that it could create a self-disclosure pressure for users that might not want to disclose their C-FTKF score to other users - a process known as "privacy unraveling" in economics [66]. Users

with good C-FTKF score, like for example users with a good credit score, might find it useful to have the dashboard in order to receive a preferential treatment from other users (and assume a position of "influencers" for example). But others with not so good C-FTKF scores may feel coerced by VoxPop to disclose their score to avoid discriminatory treatment from users. The privacy unraveling process, in the context of social media, was found when sensitive health information such as HIV status was linked to existing online identities on the geosocial hookup app Grindr [91]. We are careful into how the calibration dashboard might be shaped and we would like to utilize the four privacy unraveling limitation mechanisms proposed in [66] and further developed in [91] to solicit feedback through extensive user testing of the VoxPop beta platform.

The privacy unraveling mechanisms are: transaction cost, unverifiability of ignorance, inability to accurately infer the negative, and norms. The first limitation suggests that if the cost of disclosing is increased, the "obvious choice" of displaying one's credibility becomes less obvious, reducing stigmatizing signals from non-disclosures. As proposed in [91], a solution to this might be an introduction of premium users. We want to keep VoxPop a public and free platform and this might not work well in our case. The second limitation occurs when it is not possible to verify whether the disclosing user is aware of their C-FTKF not being disclosed. Authors in [91] propose addition of an "I don't know" HIV status, but in the case of Grindr the health status is a personal rather than an impersonal metric calculated by a platform as in the case of VoxPop. The third limitation occurs when an inability exists that inhibits negative inferences being accurately inferred around non-disclosure. A conceptual design with information grouping was proposed in [91] to allow users to mark a group of fields as undisclosed rather than each individually. This might be of some use for VoxPop where users are offered to select what elements they want into their calibration dashboard during the initial set up of their account. A possible suggestion here is the use of blurring, where for example VoxPop user might want to request their entire calibration dashboard to be blurred out to respect their privacy (and choose to which particular user they want to disclose it).

3.4 Users and Profiles

VoxPop is a platform for people from all walks of life, so we hope to attract the widest variety of individual users to the VoxPop platform, but also allow governmental organizations, support and advocacy groups, businesses, or any other groups to equally participate and voice their presence. Unlike mainstream platforms which may dissuade some users (e.g., Trump supporters leaving Twitter *en masse* after Twitter permanently banned the Trump's account) or alternative platforms that may alienate some organizations (e.g., the CDC, which does not have a Parler presence), VoxPop aims to have a diverse array of users with a set of varied perspectives to construct the discourse landscape and provide counter-argumentation. Without such openness, VoxPop would not be any more groundbreaking than Twitter or Parler and resort to addictive-turned-homophilic interaction [61]. We have outlined an example user profile in Figure 2 that may emerge on VoxPop. Before deploying a platform, it is difficult to make assumptions about what users may be attracted to and how users will choose to self-represent and express their interest. Therefore we want to avoid any presumptions about demographics or interests so the fictitious profile is only a basic user rendition on VoxPop (so are the ones in Figure11 and 12).

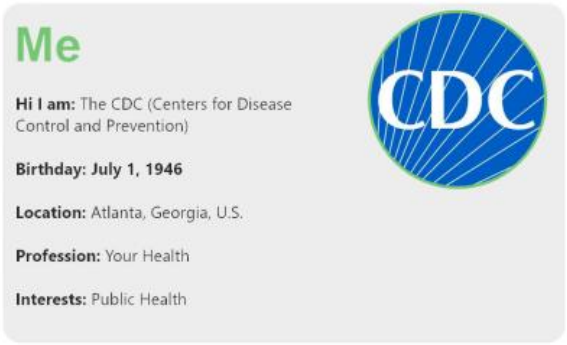

Figure 2: Example user profile on VoxPop

3.5 Social API

The minimalist visual representation is intentionally designed to emphasize the calibration dashboard as a salient feature of VoxPop. We selected to present averages and not indicate how particular users, beyond the user of the week, fare in regard to their FTKF and C-FTKF scores. The trend of analyzing political information operations, or at least dishonest behavior patterns of on social media, however, is equally geared towards analysis of individual accounts. For example, Twitter regularly publishes datasets of accounts associated with information operations campaigns in various countries [87]. The analytical engine of VoxPop is certainly capable of producing similar breakdowns and offering it to users interested in exploring the platform (instead of scraping the data, as one is forced to do with Facebook after the Cambridge Analytica scandal or Parler before the takedown). We plan to offer a social API where researchers or analysts, after approval by VoxPop administrators, can access elements of the analytical engine and create for example topical categorizations of users.

4 VOXPOP DESIGN: FRICTIONS

Frictionless UX design is motivated by a desire to increase and maintain user engagement [22]. While this allows interaction without conscious effort, it also enables users to avoid reflection or morally disengage [8]. To "undesign" the slippery slope of mindless interactions and stimulate socio-cognitive participation [90], UX designers have introduced "frictions" or inhibiting elements in the interaction to promote mindful engagement without abandoning the principles of good design [10]. In the context of "curbing" misinformation, frictions are seen as any type of interruption in the process of accessing, commenting on, or posting unverified information [46]. Twitter, for example, experimented with interstitial warning covers instead of tags as a form of a design friction that hide the entire content of a tweet with unverified information [77]. If a user wants to access such a tweet, they have to actually click on a "view" button, which is preceded by a text that explains how the tweet's content violates Twitter's "truth keeping" policy.

Such frictions increase the entry cost for participating in online discussions and prolong or disrupt the immersive nature of the scrolling through social media feeds [46]. A trade-off, therefore, must be made in order avoid users abandoning VoxPop because participation costs are too high, distraction due to frequent and often cumulative imposed frictions, or an excessive "fear-of-missing-out (FoMO)" [52]. In this section, we present each of the possible design frictions without any particular policy of appearance, order, preference, or exception rules. Instead, we will develop VoxPop's policy of friction management with usability studies to reflect our commitment to enabling users to construct the discourse contexts to the best of their preferences [90].

4.1 Sourcing Friction

In VoxPop, we have sought to utilize friction by adding a requirement sourcing of information during the posting process. Notably, when posting content that exudes high levels of faithfulness to known facts, VoxPop users will be prompted to "source" the post using a friction before the post is shown in the feed, as shown in Figure 3. Users are compelled to select if the post is their personal opinion, objective news, commentary, or satire. To increase the voice of their post, the users could add a source URL. This step signals that VoxPop trusts users with their argumentation during a discourse, but also prompts the user to switch to System 2 reflection on "why all these options?" The first step also helps the VoxPop analytical engine to calculate base FTKF and FTKF-ratings for posts, as shown in Figure 4 that help with finer social calibration.

Figure 3: The sourcing friction on VoxPop encourages users to post references, and clearly label opinion, commentary and satire before they make a post.

Of course, a possibility exists that the URL itself contains unfaithful to known facts information, has content with "expired" facts, or is a URL to a Tweet that has been soft moderated. In response, a background classification algorithm based on VoxPop's social network structure and propagation features flags such URLs and reports to the admins for further inspection based on the concept in [55] and using the analytical engine's calculations of the C-FTKF score applied to news content. The other way of gaming the sourcing friction is possible too; a user could post a URL from a known misinformation website and say something insightful about it or a user could use a URL from a "factual" source only to discredit it with a commentary unfaithful to known facts. The FTKF and C-FTKF scores do help with weeding out of such behavior, though we acknowledge that further work needs to be done to explore possible evasion strategies against the URL element of the friction.

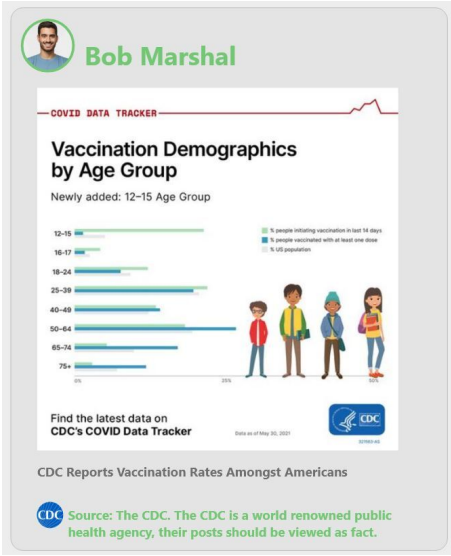

Figure 4: An example post on VoxPop, sourcing a COVID-19 data tracker from the Centers for Disease Control and Prevention (CDC) [17].

Users are allowed to skip both options in the first step by clicking the tiny "skip" option (resembling the usable security affordances used to counter phishing websites in browsers). If the user skips the "opinion" and "URL sourcing" options, in the next, more *calibrating* step, the user is asked to "supply at least couple of keywords as categories to which their post belongs for additional context". The user in this step has no option to go back to the previous step, forcing them to make a choice: abandon the post if they exude low levels of confidence in relaying "unverified" content, or yielding classification categories associated with their content that helps VoxPop's analytical engine to calculate the FTKF score of the post, which is a metric we created to reflect VoxPop's dedication to social calibration. Habitation may also happen with this friction when users learn about the option for posting content as "opinion" and opt for it even when they relay "unverified" content that will result in a low FTKF score regardless, but we address this by considering the user' C-FTKF score in Section 5).

Initially, a user that just joined VoxPop will be presented with the friction for every post. If the user achieves a high C-FTKF score, the friction will change from a choice cover to a text box that automatically determines if the content typed and posted is an "opinion" (absence of a source URL) or a "citation" (presence of a source URL).

Users could also drop to a low C-FTKF score, which will trigger the platform to bring the two-step friction back for future posts. Obviously, this adaptive interference may ultimately prolong, but not eliminate, the motivation of specific users determined to disseminate information unfaithful to known facts and content with low FTKF scores, but we believe that the posting friction will increase users' self-determination worth on a long run.

With this friction, users are offered the possibility to reflect on contextualizing their posts, instead of mindlessly replicating their stream of consciousness. Certainly, the sourcing friction is not immune to backfiring, similar to the warning labels and covers on Twitter [19]. Users could find the friction's cost too high – especially those who are highly active and very keen on posting content with various levels of faithfulness to known facts. We are aware of such a possibility and envision the sourcing friction as adaptive over time, instead of a static "undesign" feature.

4.2 Watermarking Friction

VoxPop, as indicated, offers radically re-conceptualized approach for communicating (un)faithfulness to known facts. The current warnings applied by the mainstream social media platforms are either ignored, result in a counter effect (make people believe the misinformation *more*, not *less*) [56], or cause users to feel that these platforms are biased [32] and target specific users or groups [80]. The soft moderation always includes a verbose message (interstitial covers) or a question mark to indicate the moderation (in the color scheme of the platform [74, 82], placed underneath an alleged misinformation post as a contextual tag). These are interesting design choices because warnings are usually communicated before the user is about to encounter allegedly harmful content, and they are usually in red (e.g., browser warnings for phishing). We combined this observation with the notion of "watermarking" to offer a friction that could be optionally applied to posts with low FTKF scores, as shown in Figure 5.

The watermarking friction inversely ties the percentage of the visual transparency of the watermark with the FTKF score of a given post—the more a post drifts toward unfaithfulness to known facts, the less transparent the watermark is. In Figure 5 we show an example for posts that includes images or videos, but an adaptation could be easily made for the watermark friction to appear over textual posts. The variable-transparency watermark differs from the soft moderation warnings in that it is not *after* a post and it is displayed as a red flag. We chose a "red flag" to stay true to the unambiguous signal that provokes people to stop and switch to System 2 reasoning. The red flag also allows for VoxPop to avoid including verbose content and being perceived neither as a "censor of political viewpoints" (by right-leaning users) nor as a "inconsistent and unfair moderator" (by left-leaning users) [79]. The watermark is also not *before* the post (like the Twitter interstitial covers) but is shown *together* with the post to signal an intentional distortion of the normal user experience on the platform that stays and possibly changes with the evolution of the post. This makes it hard to ignore it but could also be a reason for "backfiring" [98]. Therefore, the watermark does not have to be centrally placed and VoxPop allows for experimentation, perhaps for placement on the sides or corners.

4.3 A Moment of Mindfulness Friction

VoxPop also includes an *a posteriori* friction if a user is posting, or commenting on, large amounts of content with low FTKF scores in a short time span while consistently keeping a low C-FTKF score. We expect that there will be users with "agendas" to educate the people about the upsides or downsides of controversial issues (for one, this mission got Robert F. Kennedy Jr. banned from Instagram for spreading COVID-19 misinformation [39]). The intention of the friction is to introduce more a break and let the user take "a moment of mindfulness" for a period of time, bearing a distant resemblance with Twitter's initial time-out element in their striking system.

Unlike Twitter's time-out element, which applies a 12-hour account lock for the first "three strikes', the "moment of mindfulness" does not lock the user account but only disables the posting/commenting function for the next hour. The actual effect of the "moment of mindfulness" then takes effect because the user is offered, as part of the notification message about the friction shown in Figure 6, to browse and read posts with higher FTKF scores on

(a)

(b)

Figure 5: (a) Example watermarking friction indicating a post containing misinformation with a 50% transparency red flag watermark (b) the same post with a 25% transparency of the watermark after it generated a significant portion of other posts/comments with lower FTKF scores.

various topics, not just their particular interests (a nudge to step out of their echo chamber). The "moment of mindfulness" does not increase the effect with every other time a user takes such a moment, like in the case of Twitter, where the next time-out element for the fourth strike results in a 7-day account lock. Instead, as part of the next "moment of mindfulness", the user is asked to perform some actions as part of the "community service" to the platform.

Figure 6: "A Moment of Mindfulness" Friction: An Example

For example, a user taking a "moment of mindfulness" is notified that this is the second "moment of mindfulness" they have taken, and to restore their options for posting and commenting back, they need to help VoxPop to label posts as part of the second step in the sourcing friction. If and after the user agrees to do this and submits at least 20 data labels, VoxPop automatically enables the commenting/posting features in good faith. These labels could involve asking true/false whether the user considers a reference to an article to be misleading, or to rate how angry a post is on a scale of 1 to 5. These responses will be compared with other responses to similar questions to protect against data poisoning attacks, and the labelled datasets will be provided to the VoxPop analytical engine for integration into models. Users who abuse the system by providing bad labels can be given an additional moment of mindfulness, if the user continued their "evangelical" mission even after a good faith pass was given to them. The third and any subsequent moments of mindfulness (up to seven), will progressively increase the number of keywords/topics that the user must perform their community service on. The seventh, and last, "moment of mindfulness" will result in an automatic"factory reset" of the user profile with a negative C-FTKF score.

4.4 Suspending Friction

VoxPop, obviously, does not expect that only information and misinformation will circulate on the platform. Inevitably, there will be users that will voice hateful, racist, offensive, phobic, dehumanizing, abusive, or oppressive content, both in explicit and implicit form [60]. VoxPop administrators reserve the right to resort to a suspending friction in term of a temporal ban for posting or commenting to/with such content and user profiles if the frictions, affordances, or the analytical augmentation for constructive discourse fail to incite a reflection that will counteract any intentional effort for causing harm to other users and damage the platform (e.g. moral disengagement gone correspondingly harmful). The application of this *a posteriori* friction is informed, in addition to the input from the analytical augmentation, by the option for users to report such platform abuse. It also considers that any toxic language varies by relationship type, e.g. amongst friends on the platform or between users with no connection [69]. The suspension friction is designed to allow for revision per user's request and to leave a possibility of change of heart (of course, any repeated offenders, Sybil accounts, or any offenses that resemble a pattern after a close investigation factor in a decision for the duration of this friction) [101].

5 VOXPOP DESIGN: ANALYTICAL AUGMENTATION

Augmenting the social calibration of the discourse with analytics entails careful design of sociocognitive signals that also have utility for one's self-determination on VoxPop. Our brief treatise of "truth" provided a pragmatic genealogy between trust as an unavoidable dimension of social interaction and existence of agreed truths as fundamental ingredient or lubricant for a natural discourse. The complications of this pragmatism are elucidated when trust is defined as one's *willingness* to take a risk in a social interaction [53]. This means being "vulnerable" when participating on VoxPop, which is not a problem itself, but anonymity and moral disengagement are go-to ingredients that systematically exploit this vulnerability on the current social networks. One could propose measuring hate, rage, fear, or shame as sociocognitive signals in helping users protect this vulnerability. However, this might not be helpful because of (i) perception of the platform as being a "moral-keeper;" (ii) volatility of the emotional states and their manifestation in a discourse; (iii) highly subjective interpretation of a score indicating each of these states.

To avoid this and yet provide close to objective (again, this depends on the context), relatively stable, and neither "truth/moral-keeper" nor "free speaker" tainted sociocognitive signal, we opted to measure how much content is *faithful to known facts* and to what degree the (un)faithfulness can say about a user when is accumulated over time. Therefore, we designed the FTKF and C-FTKF scores as to build upon the social calibration characteristic for Wikipedia as a social network of editors faithful to known facts [78]. Posting/editing on Wikipedia is not highly interactive *per se* but is nonetheless dedicated to bring the known facts to the fore. If we utilize a similar social calibration approach and signal the measurements of the said process, we believe it could help users participate in a discourse *whilst* willing to take risks (e.g. being vulnerable). Here, discourse could manifest as dialog, debate, but also a disagreement, and perhaps include forms of provocation. We kept away from measuring "truth," even if empowered by the "wisdom of (selected) social media crowds [11, 107], because seems hard to scale in practice, could be error-prone and biased (e.g. selective consideration of facts), and slow to resolve conflict.

On existing social media platforms, users orient their success and engagement through metrics on the account level (followers/following counts) and metrics on the post level (number of people that viewed, liked, commented, or shared/retweeted a post) [76]. VoxPop retains these metrics but also introduces and exposes users to an account level (C-FTKF score) and post level (FTFK score).

The idea behind this is the tendency to self-regulate behavior based on scores, for example, the FICO credit score, Grade Point Average (GPA), ride-sharing score or five-start reviews, to name a few. By presenting these scores to the users in their dashboards, VoxPop provides a glimpse into how they fare at a particular point of time among their friends or the platform in general.

The use of the dashboard, and public posting of a user's metrics create a gamification context. As a strategy, gamification has been used to inform and change users' behavior [47], for example complicated technical concepts [70] or simple physical exercise [86]. While there is a risk that sharing this information publicly could lead to this gamification being a part of a "dark pattern" where the gamification leads to users feeling like their privacy is being breached [100]. In the context of VoxPop, the dashboard features are aggregating an individual user's content which they have previously shared, additionally, it may increase an individual user's self-awareness regarding the quality of the content that they post. The goal is to empower the users to consider their self-representation while participating in the discourse on VoxPop instead of only chasing followers, likes, or shares. The ultimate idea, therefore, is to incentivize users to participate in the "game of known facts" with a socially-outward attitude instead of seeking inward-focused individual gratifications (e.g. authors in [86] show that socially-outward fitness apps are better in helping consumers sustain their efforts in physical activity than apps focused on individual fun as gratification for exercise).

5.1 Faithfulness-To-Known-Facts (FTKF) Score

We are currently working on identifying an ideal method for identifying how "faithful" to known facts a post is quickly and efficiently based on the reputations of the person making the post, and the source of that post's content. These can be used to identify drops in ratings for sources, e.g. posts that reference a specific source are suddenly being used to advance fallacious claims, or with specific users, e.g. a user who normally verifies their sources, has started sharing a large amount of content from poorly rated sources. These scores are dependent on users' ratings, so we assume "data sparsity" where the majority of users only contribute a small number of content faithful to known facts, as opposed to lopsided cases where misinformation propagates much faster than verified information [107]. This dependency on user ratings is similar to the assumptions baked into Twitter's Birdwatch program [88].

We set to determine the *FTKF* score, assigned to each post, using the analytical information collected from the sourcing friction. VoxPop also presents yes/no/maybe radio buttons underneath a post to a random selection of users to solicit input on the validity of the posts. These ratings are aggregated as a "base score," and could be calculated based on the reactions to a single post, or a sample of posts written by a user, based on the rate of people who vote that a source is faithful to known facts r, with "maybe" votes counting against a user. These ratings will be based on the number of people to rate them n, and sample size hyper-parameter ss, which can be manually set by the VoxPop admins, where larger values mean that more people must vote to achieve a score of 5. The values are scaled to be between 0 and 5, to mimic the common 5-star rating system used to assess quality in online settings.

$$FTKF = 5 \times \min\left(1, \frac{n}{ss}\right) \times r$$

While this is a very rudimentary approach to calculate a FTKF score, it could be assigned and updated in near real-time, to reflect the developments in opinions, reports, and emerging contradictions in the discourse (we must note that the score is a subject to much more in-depth analysis and marked improvement in the next stages of the VoxPop's development). Long-term, additional consideration could be given to using natural language processing techniques to potentially try to identify linguistic patterns in the posts' text that are correlated with labels derived from the FTKF and C-FTKF scores. Deviations between predicted FTKF scores and the FTKF score labels could serve as another layer of detecting deviation from what is known to be rooted in facts.

Another possible behavior is where users deliberately downvote true stories they do not like, e.g. an article from a reputable journalist that accuses a popular celebrity of inappropriate conduct. This provocative voting would enable the fanatical followers of this celebrity to, as far as VoxPop is concerned, decide on a set of facts that are not in line with reality. In order to help prevent this and other manipulation of FTKF scores, as an added layer of protection, the ability to vote on the faithfulness to known facts of specific posts will be random. The randomness makes it more challenging for users to seek out, and artificially raise the FTKF score for specific news sources or false narratives. There may also be situations where there are multiple concurrent narratives, any of which are equally likely to be true on a topic, leading to different people rating content different ways based on simple disagreements. In this situation, users are not voting based on faithfulness to known facts, they are voting based on speculation, and simply abstaining from voting is the best course of action that the user could take. As these situations start to arise, we will have data that can be used to help identify how users behave in these situations, and to update our platform accordingly.

The FTKF score is thus not immune to manipulation. Within VoxPop, echo-chambers can start to form where large numbers of people may rate posts from websites like *Infowars* or *Lindell TV* as truthful because that group has come to the consensus that those articles are legitimate. In these situations, the FTKF may still be able to help, because analysts can calculate the FTKF for arbitrary groupings of posts, such as, the ratings of posts from specific source URLs, and these other FTKF calculations can in turn be used to identify whether a specific cluster of users may be rating a source significantly higher than users across the rest of VoxPop rate that source. This is still, of course, *a priori* speculation, and the efficacy of the FTKF in identifying "unfaithfulness" to known facts remains to be seen.

5.2 Cumulative FTKF (C-FTKF) Score

While the FTKF score helps with basic calibration of the discourse on VoxPop, it is hard to expect that all users will act honestly and achieve an equilibrium in being faithful to known facts (facts themselves could change over time, for one). There will be users with dishonest intentions that could take advantage of the constructive discourse to post misinformation with regard to reaching the most users in a critical window or high-truth relativity (it might

be overused as an example, but the early promulgated COVID-19 misinformation social media content resembles this scenario [40]). We might be able to avoid this situation by having a cumulative FTKF (C-FTKF) score to be used on a user level. This could behave like Reddit's reputation score, where more time on the platform means it is possible for the user to continuously increase their C-FTKF scores. Additionally, in order to protect our metrics from bot attacks tanking the ratings, we will consider using a graph-based approach like SybilLimit [103] to identify inorganic interactions between members of different nodes, or the Advogato trust metric that measures the reliability of individual members [50]. Factoring the number of posts the user generates and their FTKF score one could do either (1) exponential or (2) ratio accumulation as shown in Figure 7. We favor the first option because it could be augmented with the standard metrics of post engagement and user interaction in fine-tuning the gradient of the exponential accumulation over time and avoid reputation stagnation.

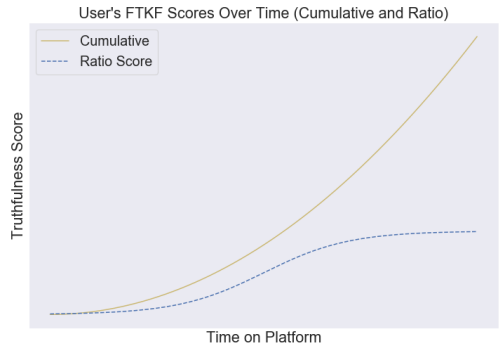

Figure 7: A graph comparing cumulative and non-cumulative approaches to FTKF scores. Over time, a ratio score provides less motivation for a user to continue posting content faithful to known facts as their account matures, while an exponential cumulative score based on user interactions provides even more motivation for a mature account to continue being faithful to known facts.

5.3 Disinforming Posts and Followers Count

There is a high probability that the above calibration metrics might not be sufficient for overcoming emerging challenges in detecting and handling sociality with dishonest intentions on VoxPop (which in this section we treat as "disinformation" to emphasize the intention of spreading unverified and inaccurate information toward fulfilling nefarious agendas [104]). The changes to user behavior may make it more difficult to use more traditional classification techniques for identifying content with questionable provenance and we gear our analytical frictions more toward sociocognitive "signaling" to users about the perils of dishonest behavior on VoxPop rather than downright taking punitive actions (or avoiding, as is the case with the alternative platforms). We chose to do this in order to stay true to the commitment of social calibration but also to avoid incidents steaming from *en masse* classification like the one where Twitter automatically banned all accounts that posted or commented using the word "Memphis" [37].

One challenge is that the focus on explicit disinformation removal from the platform may result in the perception of legitimacy of information that remains on the platform that has a high probability of being faithful to known facts. This means that people who are looking to start a disinformation campaign might be drawn to VoxPop because of its potential to add credibility to their voice. To help avoid such dishonest behavior, we did some rudimentary modeling of user incentives, and it seems that at any point a user hoping to spread disinformation has two choices: (1) they can post a high FTKF score post, which can boost their C-FTKF score and help them to gain more followers (probably using the option to add an URL through the sourcing friction); and (2) they can post disinformation (supplying benign keywords, if, for example, they target a controversial issue), which, if identified can severely hurt their C-FTKF scores. By selecting the second option, users risk losing followers, though, and reduce the likelihood of future misleading or harmful posts from being perceived as legitimate.

The value of these different actions will gradually change, however. If a political operative trying to get their preferred candidate elected, then it is worth more to them to post to a smaller following before an election than to a larger following after the election. Also, if a user is primarily posting high FTKF score content to build their following, eventually they will start to reach a limit where they are not getting as many new followers or increasing their brand as much for each post with a high FTKF score. This could be, for example, because they are posting primarily liberal content in Michigan, and there are a finite number of liberals from Michigan on the platform who would be interested in following the user. The corollary is that there will be specific points where it is especially advantageous for users to post misleading content. One way we anticipate that we will be able to identify these points is to model user engagement across their posts, perhaps modeling individual users using an ordinary least squares model with a `sigmoidal` term, and take the derivative of their curve, which will help to inform us where in the journey an individual is, as shown in Figure 8.

5.4 Follower Dropout

While ideally a user's total number of followers would decrease after each post with a low FTKF score, it seems like they might be able to quickly rebuild their total number of followers and continue to grow long term, meaning that they may be able to routinely exploit the sourcing friction and keep post disinformation. Users unfollowing the user they perceive as dishonest and negatively contributing to the discourse depends on the users being able to quickly identify that a post is disinformation (for which the FTKF score provides indication of). In any case, this activity of organic unfollowing, as we envision it, is shown in Figure 9. Of course, the outlook for rebuilding user's reputation will depend on the disinformation topic, the user's C-FTKF score, and factors associated with unfollowing in general (e.g political activity or emotional involvement) [99].

One way to use follower count to stop adversaries from using VoxPop for misinformation dissemination is, when a user makes a post that is demonstrably disinformation and will certainly result in an extremely low FTKF score, to have a percentage of their followers automatically "unfollow" them. This strategy would mean that after

a user posts disinformation, they have to rebuild their follower count before posting another piece of disinformation, or else they lose another percentage of their followers as shown in Figure 10. This presents a user who is planning to post disinformation with a dilemma in which they have to make a very explicit decision regarding whether they would like to risk losing a percentage of their followers.

Bad actors and dishonest users know well that disinformation without followers is nothing [105], which could result in a migration to less regulated platforms. However, VoxPop does not signal that it targets any user susceptible to a particular piece of disinformation, but that there is a "price to pay" if a user wants to play with disinformation (this friction also challenges trolls to adopt a nuanced game-theoretic approach in participation on VoxPop instead of a simple "make or break" strategy). VoxPop allows users who unfollow someone to have the option to re-follow them. This would mean that a user does not lose their ability to follow whomever they want but would make it more work for a user to follow someone who is repeatedly posting disinformation than to unfollow people who are posting disinformation.

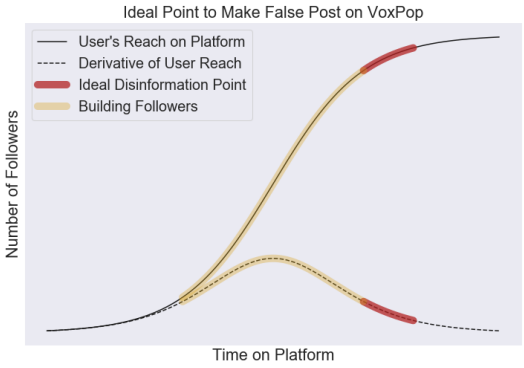

Figure 8: A graph showing a possible method for detecting when a user is more heavily incentivized to post disinformation—after they are no longer gaining substantially more followers per post], which is indicated in the decrease in the first derivative of the `sigmoidal` pattern that we anticipate a user's follower count as they remain on the platform to follow.

5.5 Guilty Until Proven Innocent?

One challenge is to determine whether it is better to start a user with a very low score that they can increase by posting posts faithful to known facts, or whether their initial score ought to be high when they are on the platform until they reduce their scores. Starting users with a low score may dissuade people from following them when they already do not have very many followers. This means that it will take a user a lot more effort to build their C-FTKF score. Additionally, this may skew platform-wide metrics based on the number of people who start out with very low scores, giving a false baseline that users compare themselves to. Starting users with a high score might mean taking on additional risk for the platform due to the new user's relatively few followers. This may, however, result in users' performance metrics gradually regressing to the mean, and the user may find that this artificial negative progress disincentivizes them from trying to post content faithful to known facts to boost their metrics. One strategy could be adapted from Twitter's Birdwatch, which, before collecting at least five ratings, says that a tweet "Needs More Ratings" [88]. This could be used generally in both individuals' performance metrics, as well as in the FTKF scores associated with each post.

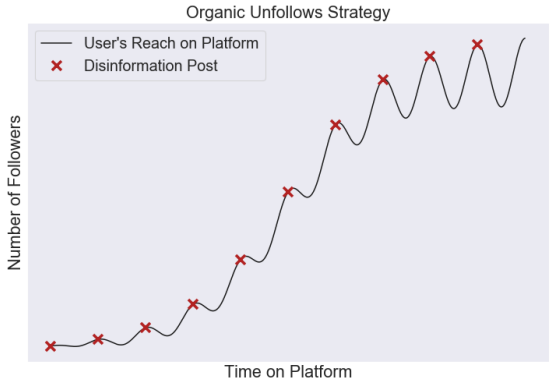

Figure 9: A graph showing users organically unfollowing someone's posts after they are marked as being misinformative with a low FTKF score. Notice that the user is able to rebuild their following relatively quickly, and can continue posting disinformation and having it reach a wide audience.

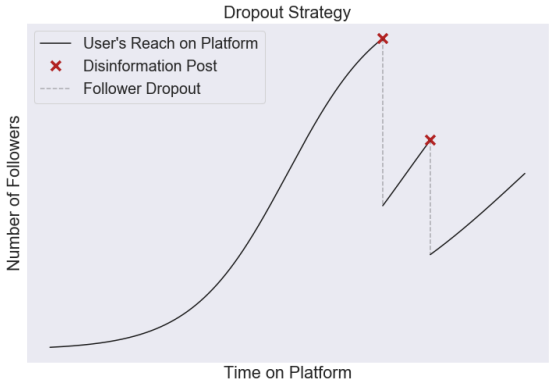

Figure 10: A graph showing the dropout strategy, where a percentage of a user's followers are automatically removed after making a post with a low FTKF score. Note that after each decrease, the total threshold of people willing to follow them decreases.

6 VOXPOP CHALLENGES

6.1 VoxPop Ethics

The public discourse on social media lends itself to a series of ethical concerns regarding *displaying content* when it is harmful and *hiding content* which could be construed as censorship. Due to these issues, platforms have to walk a fine line between removing harmful content while still protecting freedom of speech. The ethical tension, then, arises between the need to be perceived as a "truth keeper" (of course, as long as that is desirable from a public relation perspective) and the need for remaining profitable or true to the cause—alternatively, between the "free speaker" proclamation and the over-indexing on "influencers" to speak for many [7]. Perhaps it is too early to say, but we envision this tension exacerbating where the platforms are deemed as "big social media" (akin to "big Pharma", "big Tobacco", etc).

The initial "crisis of misinformation" was an opportunity capitalized on by all platforms, but the unfaithfulness to known facts is unpredictable and increasingly becomes a burden for profit and attention (mainstream) [97] or burden of differentiation (alternative) [61]. True, social media platforms could not be universally liked, or perceived as ideal, and we certainly do not expect that to be the case with VoxPop. The experimental ideas put forth within VoxPop, could place the platform in either the "truth-keeper" or "free-speaker" camps despite our initial idea to assume no profitability or "socially agreeable context" approach in the design. For one, VoxPop could create circumstances of "implied faithfulness" to known facts with the scores, affordances, and frictions, which might not perfectly overlap with the real, actual faithfulness nor the actual, real facts available to participants elsewhere. [64].

VoxPop, as a social media platform, has a public function in that it (1) *facilitates* public participation in art, politics, and culture; (2) *organizes* public conversation so people can easily find and communicate with each other; and (3) *curates* public opinion through feeds, moderation, and regulating speed of content propagation [7]. We adhere to these principles and posit that VoxPop enables democratic participation in the formation of public opinion and that social calibration and freedom of speech support the growth and spread of knowledge. In fact, VoxPop insists on the democratic value of listening to the other and rests on the idea that unfaithfulness to known facts is a form of communicative action rather than an epistemological statement about reality, which we must learn to deal with rather than try to remove [2]. VoxPop's paradigm aligns with the postulation that "political truth is not discovered and then told but generated through acts and modes of telling" and attempts to capture the effect that follows (i.e., that ways of speaking truth change, often quite dramatically, in response to emergent technologies, genres, and vocabularies of mediation) [20].

VoxPop draws, in part, from the agonistic pluralism model of democratic discourse grounded in productive conflict or contest where democracy is cast as an endeavour of fervent competition and struggle among competing ideals, values and beliefs [25]. Extending the treatise of design for politics and political design within the agonistic pluralism, VoxPop could be seen as platform focused on improving structures and mechanisms that enable self-governing of social media discourse, and as such, the VoxPop's design to be seen as design for politics. At the same time, VoxPop's design could be seen as political design because the experimental nature of VoxPop allows for critical investigation of the "crisis of misinformation" issue and raise questions concerning the conditions of this issue. Time will tell whether and how the VoxPop experiment will live to this envisioned paradigm new social networking platform under the sun. At least VoxPop dares to act instead of only generating yet another data-backed description of this so-called crisis and concluding with a "call for action".

6.2 Bias

VoxPop's idea for using the FTKF and C-FTKF scores for social calibration, though noble and promising in nature, could have unintended consequences on the actual democratic opportunities for participation in the public discourse. Algorithms, contextualized with ambiguous rules and often exploited with carefully crafted content, could be perceived as biased and VoxPop recognizes this as a serious challenge for the platform's adoption [80]. Although VoxPop tries to combine sourcing toward forming an implicit consensus on baseline elements in any discourse, there are many topics that have equally relevant but countering versions for them [31]. Welcoming diverse perspectives is one of the core values of VoxPop but we are not excluding the possibility that users might interpret the frictions and the social calibration as actions for their censorship, suspension, or shadowbanning.

Studies have shown that the most common theories offered by users about moderation on social media are that (1) their content was flagged by another user; (2) the platform is politically biased against them; or (3) that the platform does not uphold norms of free speech under U.S. law [42, 93]. Given that VoxPop rests on the idea of social calibration, the theory that might come first to practice is the first one, with users feeling that their C-FTKF scores do not reflect their self-perceived reputation or the reputation they enjoy in their real-world circles. We are aware of the threat of "moderation opacity" [71], therefore we allow for users to appeal and participate in further improvements and adjustments of the calibration metrics as well as frictions. VoxPop takes the responsibility to inform users about the principles of calibration and frictions on the platform in order to prevent users from wondering where they went wrong in case they post or comment information they believe deserves a higher (or lower) truthfulness score than they expect.

In regard to UX design bias, the description of VoxPop frictions in the above sections allows for the participation of well-able users, but VoxPop will not go live or in any advanced testing without an inclusive platform design. VoxPop adheres to the Accessible User Experience model (AUX) and counts affected users as design collaborators with special knowledge about disabled bodies that might offer additional and innovative affordances, frictions, and metrics for social calibrations for all users [58]. VoxPop is also envisioned, in this proto-version, to serve only in English and for U.S.-centric public discourse, and we are determined to work toward eliminating the preliminary cultural and linguistic bias in order to bring the platform to a wider user base. We cannot say for sure that VoxPop's discourse will be so popular that it will result in anxiety effects such as FoMO but we could envision scenarios in which the VoxPop features might act as FoMO persuasion triggers [1, 94]. For example, users might develop a fear of dropping below a certain threshold of their C-FTKF score or fear missing the ability to

retain followers. These problems certainly push VoxPop to think of adaptations of counter-FoMO or discontinuous adoption frictions in later versions and we are determined to work on them in our future research.

It is to be expected that there are to be users with preferences for safety of their uncritical bubble than to be exposed to social calibration given that a critical element for user acceptance is the proclivity for flow experience when using a social media platform [49]. There is a chance that VoxPop's UX frictions could be perceived as breakpoints in the flow and such users might abandon VoxPop altogether. Even if users assume the UX frictions as part of the flow, they might find them reasons to mistrust the analytical augmentation algorithms of VoxPop, i.e. running the risk to be perceived as "dark patterns" [12, 26]. To minimize this risk, VoxPop provides transparency into the inner workings of the platform and welcomes dialogues with users aiming to minimize and eliminate perception of coerciveness, deception, and algorithmic strong-arming.

6.3 Antisocial Behaviors

One of the main goals of the VoxPop is to harness the power of the community to create norms around problematic and abusive behaviors. While alternative platforms might altogether "avoid" dealing with it, mainstream platforms fall short to offer a unified and agreed definition of what constitutes antisocial behavior [60]. VoxPop's terms of service clearly describe the manifestations of antisocial behaviors that have no room on VoxPop, adopted from the *Field Manual of Online Abuse* [62]. We incorporated the suspension friction as a mean throughout in which VoxPop moderators enforce the terms of service. Mainstream platforms, in addition to moderation, include affordances such as muting, blocking, or reporting offensive users [43]. Third-party applications go a step further and offer blocklists, allowing users to quickly block all accounts on a community-curated or algorithmically generated list of block-worthy accounts (akin to "killfiles" in Usenet) [33].

The proto-design of VoxPop allows for users to report problematic behavior but we left the implementation of the mute, block, or blocklist features for later versions in order to better inform our design by observing dishonest participation on the platform. If, upon revision, content undoubtedly meets any of the definitions in [62], VoxPop administrators reserve the right to apply the suspending friction in various durations with the option for appeal by the users. The decision, and a later application of users' account metrics monitoring, is also informed by the tactics used by online abusers and harassers such as brigading, concern trolling, dogpiling, dog whistling, doxxing, identity deception, sealioning, subtle threatening, swarming, and swatting [43]. These tactics are important for consideration because a single report might not constitute an antisocial behavior but considered in the context of the discourse and with other potentially unreported users might be deemed as harmful to a user, nonetheless.

We anyhow, in parallel, ideate on how to include the mute, block, and blocklist affordances. For example, we envision the "mute" affordance as a filter where users can set a custom minimum and maximum threshold of the FTKF score for a selected period of time to see only posts that match this rule. The filter will also allow for defining similar rules about the C-FTKF score in order to help users distance from or avoid, temporarily, those who perceive other users as abusive. Because on Twitter a muted user is not notified that they are muted, and they may continue posting to the user who muted them without realizing the receiver cannot see their posts, we have an idea to include an additional metric called "decibels" which indicates to a user the how loud they are "heard" in their last post, among their friends, among the friends-of-their-friends, region, and the entire platform. A user seeing a decrease in decibels, we hope, might reflect on their self-determination and contemplate why their voice does not carry of late.

Blocking a user on other social media platforms prevents that user from viewing the blocker's posts or sending direct messages to the blocker. The blocklists extend this functionality to block users *en masse*. While VoxPop sees a value in experimenting with blocking users/lists, we are wary of negative consequences such as building non-transparent echo chambers or false positive blocked users. A variant of the "block" feature could be enabled for limited use where the blocker is offered a type of friction that they deem the most appropriate to be presented to the "to-be-blocked" user(s): enforced sourcing friction (only option to post and comment with validated sources) to the blocker or a varying moment of mindfulness friction (1 hour to 1 week) with blocker-selected community service. Such a block feature can be combined with the mute feature, but it will be important for the VoxPop observe, at least in the beginning, the nefarious use of both features to ensure a balancing calibration of the discourse while maintaining the platform's openness.

There are a few possible options for methods for algorithmically breaking down the sides of an echo chamber. For example, Garimella et al. used graph-based model of user interactions on Twitter to successfully differentiate Twitter users who comprise different sides of a few different debates [31]. One of the methods they found to be the most successful was a method called the *Random Walk Controversy*, which uses the authority of certain users in certain discussions to measure the existence of debates [31]. If the VoxPop model cold identify debates, and which side of a debate the user is on, it could possibly provide a user in a debate with additional content from users on the other side of the debate. Here, we have to be careful to distinguish between debates, disagreement, and provocation in order not to appear as taking a side and stifling various types of discourses. Another approach might be to use a recommender-based system. By carefully balancing of the recommendation system's exploration phase–where the user is exposed to a variety of different content to see what content a user engages with–and its exploitation phase–where the recommendation system shows the user content that it most expects that the user will engage with [85]. Various algorithms differ on balances of these phases, with some algorithms tending to use less exploration than others. One option would be to use an algorithm with an explicit exploration hyperparameter [85].

6.4 Coordinated Manipulation and Social Bots

VoxPop is not immune to the presence of state-sponsored groups or bots, and this presents a challenge as it may be impossible to eliminate their presence [28]. Since the beginning of social media, these groups have attempted to weaponize information, and they constantly evolve to evade the detection of platforms who are always behind playing catch-up [23]. But rather than working frantically

to remove bot content, VoxPop welcomes the presence of it in order to observe what dishonest patterns of discourse will emerge as a result of bot/semi-bot activity [4]. Demystifying the bot activity on VoxPop is an essential step in evolving the platform given that bots, even in the presence of automation software, must learn and adapt to the novel affordances, frictions, and analytical augmentation. Friction-wise, VoxPop could offer a variant of "CAPTCHA" that pops up on the user's screen, asking them to tag the posts they believe are "misinformation" (or verified information) either taken from VoxPop's feed or borrowed from other social media platforms. We recognize that this friction might also spill over to actual users, and that in and of itself, it is also in our interest to do user testing and further research on.

6.5 Crowd-Sourced Calibration Manipulation

In 2006, when *The Colbert Report* host Stephen Colbert suggested his fans change the "Elephant" *Wikipedia* article to claim that the number of elephants had tripled over the last six months, users started rapidly defacing those articles, and other articles related to elephants. To combat this nuisance, *Wikipedia* administrators had to change the status of all of these articles to "semi-protected" to prevent further abuse by non-registered, non-reputable users. Likewise, external leadership and have the potential to encourage users to make illegitimate rankings on articles to give the illusion of faithfulness to known facts, especially users with status and means for information dissemination on other platforms.

If Alex Jones of *Infowars* elected to wage an information war, for example, he could encourage his viewers to help him spread his disinformation by simply rating his posts as being highly accurate and truthful. This behavior is difficult to curb and has the potential to derail lots of other aspects of VoxPop. One possibility is to use users' own metrics in the weight of their ratings of other users, but in the above example, Alex Jones could go just a little bit further to tell his viewers to share lots of content by the mainstream media, abuse the sourcing friction, and use those boosted metrics to then promote the content by Alex Jones. Ultimately, we have not yet identified a strong strategy for preventing this abuse in the future, but we might have to wait and observe the emergence of such nefarious patterns of truth manipulation to tailor VoxPop's friction/analytical response.

7 VOXPOP FUTURE

This paper conveys the first blueprint of our design ideation of VoxPop. The experimental transition of the platform is planned for later stages once we solicit feedback from designers, security and privacy professionals, software engineers, sociologists, information operations experts, accessibility advocates, and most importantly, potential future users. Because VoxPop is "for the people, by the people," rich personas based on complex behavioral and demographic traits will be drawn to inform the design of evolving and diverse voices on the platform [75]. We created VoxPop not only as a natural progression towards the next generation social networks noticing a trend of "us versus them" division among the platforms, but also to address a lack of actionable and inclusive usable security that addresses murky relationship with facts beyond banking on punitive moderation.

A similar need was recognized in other works, albeit very topical and with a different set of participatory incentives, such as the "Reflect!" platform supporting constructive argumentation among students in solving wicked problems [38] or the "ProSocial Design" network [67]. Similar to VoxPop, the "Reflect!" platform exposes users to contrasting, controversial, and varying arguments in order to address a common problem. As VoxPop, "ProSocial Design" network also realized the divisiveness of the current social media landscape and provides a set of UX design "interventions" envisioned to produce quantifiable pro-social outcomes in a constructive discourse. For example, ProSocial Design has evolved versions of the interstitial warning covers from Twitter to help with "inoculation against misinformation", and a user-enabled headline rating akin to the Birdwatch program to help "reduce sharing of misinformation".

VoxPop does not have targets for "reducing the share of misinformation" per se, but more so of "allowing for reflectivity when engaging in constructive discourse". Nor do we strictly aim to "inoculate against misinformation." Anti-inoculation exists, and perhaps will continue to exist, but the "misinformation virus" we are afraid of exhibits more parasitic and fast morphing properties that will require continuous changes and adaptations in the inoculations. VoxPop instead is focused on exploring how one might develop a natural (if not herd) immunity in presence of information unfaithful to known facts by day-in-day-out participation in a discourse that contains information with variable "virility" and "propagation".

VoxPop's affordances for materializing agonistic pluralism in either of the design for politics or the political design variants has to deal with accumulation of expectations about user heuristics, cognitive biases, and respectively behavioral tendencies to achieve successful acceptance/adoption among users as a natural progression of online sociality. Whether VoxPop will achieve such a critical mass is our next research challenge in which we set to do human-subject studies where users are given limited access to particular features to test, report, and participate in an interview with us as designers to share their impressions. We plan to use semi-structured questionnaires where we offer the participants to first give their feedback on a feature (e.g. the watermarking friction), provide them with an alternative handling on other social media sites (e.g. Twitter warnings) and ask for their preferences (alliteratively exposing two groups of users to both features and compare their feedback).

Based on this pilot study, a refined VoxPop 0.1 (beta) will be released for controlled used with a bigger population for testing over a longer period (e.g. several months) and solicit feedback. We don't know if in the end VoxPop as an overall experiment will work and may very well be an entire failure. But, even if only one feature succeeds and is adopted by any other social media platform, we will consider it a success, nonetheless, because our vision of is for VoxPop to open a turf for battle with the ultimate prize of knowledge circulated on the platform thorough innovating affordances, frictions, and user-tailored analytical augmentations.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Volker Roth and Dr. Kevin Gallagher for the dedicated and sincere help shepherding our paper. We would like to thank Mark White for the support in designing the visuals of VoxPop to convey our ideas in the best possible way.

REFERENCES

[1] Aarif Alutaybi, Emily Arden-Close, John McAlaney, Angelos Stefanidis, Keith Phalp, and Raian Ali. 2019. How Can Social Networks Design Trigger Fear of Missing Out?. In *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*. 3758–3765. https://doi.org/10.1109/SMC.2019.8914672

[2] Jack Andersen and Sille Obelitz Søe. 2020. Communicative actions we live by: The problem with fact-checking, tagging or flagging fake news – the case of Facebook. *European Journal of Communication* 35, 2 (2020), 126–139. https://doi.org/10.1177/0267323119894489

[3] Rudolf Arnheim. 1957. *Art and visual perception: A psychology of the creative eye.* Univ of California Press.

[4] Dennis Assenmacher, Lena Clever, Lena Frischlich, Thorsten Quandt, Heike Trautmann, and Christian Grimme. 2020. Demystifying Social Bots: On the Intelligence of Automated Social Media Actors. *Social Media + Society* 6, 3 (2020), 2056305120939264. https://doi.org/10.1177/2056305120939264 arXiv:https://doi.org/10.1177/2056305120939264

[5] Bence Bago, David G Rand, and Gordon Pennycook. 2020. Fake news, fast and slow: Deliberation reduces belief in false (but not true) news headlines. *Journal of experimental psychology: general* (2020).

[6] Christopher A. Bail, Lisa P. Argyle, Taylor W. Brown, John P. Bumpus, Haohan Chen, M. B. Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky. 2018. Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences* 115, 37 (2018), 9216–9221. https://doi.org/10.1073/pnas.1804840115 arXiv:https://www.pnas.org/content/115/37/9216.full.pdf

[7] Jack M Balkin. 2020. How to regulate (and not regulate) social media. *Knight Institute Occasional Paper Series* 1 (2020).

[8] Albert Bandura. 2016. *Moral disengagement: How people do harm and live with themselves.* Worth publishers.

[9] Robert E Bartholomew and Robert W Baloh. 2020. Challenging the diagnosis of 'Havana Syndrome' as a novel clinical entity. *Journal of the Royal Society of Medicine* 113, 1 (2020), 7–11.

[10] Steve Benford, Chris Greenhalgh, Gabriella Giannachi, Brendan Walker, Joe Marshall, and Tom Rodden. 2012. Uncomfortable Interactions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Austin, Texas, USA) *(CHI '12)*. Association for Computing Machinery, New York, NY, USA, 2005–2014. https://doi.org/10.1145/2207676.2208347

[11] Md Momen Bhuiyan, Amy X. Zhang, Connie Moon Sehat, and Tanushree Mitra. 2020. Investigating Differences in Crowdsourced News Credibility Assessment: Raters, Tasks, and Expert Criteria. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2, Article 93 (Oct. 2020), 26 pages. https://doi.org/10.1145/3415164

[12] Kerstin Bongard-Blanchy, Arianna Rossi, Salvador Rivas, Sophie Doublet, Vincent Koenig, and Gabriele Lenzini. 2021. "I am Definitely Manipulated, Even When I am Aware of it. It's Ridiculous!" - Dark Patterns from the End-User Perspective. *Designing Interactive Systems Conference 2021* (Jun 2021). https://doi.org/10.1145/3461778.3462086

[13] Danah M. Boyd and Nicole B. Ellison. 2007. Social Network Sites: Definition, History, and Scholarship. *Journal of Computer-Mediated Communication* 13, 1 (2007), 210–230. https://doi.org/10.1111/j.1083-6101.2007.00393.x

[14] Rupert Brown and Samuel Pehrson. 2019. *Group processes: Dynamics within and between groups.* John Wiley & Sons.

[15] Fink Brunton. 2015. *Spam: A Shadow History of the Internet.* MIT Press. https://books.google.com/books?id=tJ0jEAAAQBAJ

[16] Manuel Castells. 2011. *The rise of the network society.* Vol. 12. John Wiley & Sons.

[17] Centers for Disease Control. 2021. COVID-19 Data Tracker. https://covid.cdc.gov/covid-data-tracker/#datatracker-home

[18] Matteo Cinelli, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. 2021. The echo chamber effect on social media. *Proceedings of the National Academy of Sciences* 118, 9 (2021). https://doi.org/10.1073/pnas.2023301118 arXiv:https://www.pnas.org/content/118/9/e2023301118.full.pdf

[19] Katherine Clayton, Spencer Blair, Jonathan A Busam, Samuel Forstner, John Glance, Guy Green, Anna Kawata, Akhila Kovvuri, Jonathan Martin, Evan Morgan, et al. 2019. Real solutions for fake news? Measuring the effectiveness of general warnings and fact-check tags in reducing belief in false stories on social media. *Political Behavior* (2019), 1–23.

[20] S Coleman. 2018. The elusiveness of political truth: From the conceit of objectivity to intersubjective judgement. *European Journal of Communication* 33, 2 (2018), 157–171. https://doi.org/10.1177/0267323118760319

[21] A.B. Cox and C.M. Rodríguez. 2020. *The President and Immigration Law.* Oxford University Press.

[22] Anna L. Cox, Sandy J.J. Gould, Marta E. Cecchinato, Ioanna Iacovides, and Ian Renfree. 2016. Design Frictions for Mindful Interactions: The Case for Microboundaries. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI EA '16)*. Association for Computing Machinery, 1389–1397.

[23] 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 *Proceedings of the 26th International Conference on World Wide Web Companion* (Perth, Australia) *(WWW '17 Companion)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 963–972. https://doi.org/10.1145/3041021.3055135

[24] Michael A. DeVito, Jeremy Birnholtz, and Jeffery T. Hancock. 2017. Platforms, People, and Perception: Using Affordances to Understand Self-Presentation on Social Media. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Portland, Oregon, USA) *(CSCW '17)*. Association for Computing Machinery, New York, NY, USA, 740–754. https://doi.org/10.1145/2998181.2998192

[25] Carl DiSalvo. 2010. Design, democracy and agonistic pluralism. In *Design and Complexity - DRS International Conference*. 1–10.

[26] Verena Distler, Gabriele Lenzini, Carine Lallemand, and Vincent Koenig. 2020. The Framework of Security-Enhancing Friction: How UX Can Help Users Behave More Securely. In *New Security Paradigms Workshop 2020* (Online, USA) *(NSPW '20)*. Association for Computing Machinery, New York, NY, USA, 45–58. https://doi.org/10.1145/3442167.3442173

[27] Andrew J. Elliot and Markus A. Maier. 2014. Color Psychology: Effects of Perceiving Color on Psychological Functioning in Humans. https://www-annualreviews-org.ezproxy.depaul.edu/doi/10.1146/annurev-psych-010213-115035

[28] Emilio Ferrara, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini. 2016. The Rise of Social Bots. *Commun. ACM* 59, 7 (June 2016), 96–104. https://doi.org/10.1145/2818717

[29] Pabini Gabriel-Petit. 2007. Applying Color Theory to Digital Displays. https://www.uxmatters.com/mt/archives/2007/01/applying-color-theory-to-digital-displays.php

[30] H.G. Gadamer. 2013. *Truth and Method.* Bloomsbury Publishing.

[31] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2018. Quantifying Controversy on Social Media. *Trans. Soc. Comput.* 1, 1, Article 3 (Jan. 2018), 27 pages. https://doi.org/10.1145/3140565

[32] Christine Geeng, Tiona Francisco, Jevin West, and Franziska Roesner. 2020. Social Media COVID-19 Misinformation Interventions Viewed Positively, But Have Limited Impact. arXiv:2012.11055 [cs.CY]

[33] R. Stuart Geiger. 2016. Bot-based collective blocklists in Twitter: the counter-public moderation of harassment in a networked public space. *Information, Communication & Society* 19, 6 (2016), 787–803. https://doi.org/10.1080/1369118X.2016.1153700

[34] Nabeel Gillani, Ann Yuan, Martin Saveski, Soroush Vosoughi, and Deb Roy. 2018. Me, My Echo Chamber, and I: Introspection on Social Media Polarization *(WWW '18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 823–831. https://doi.org/10.1145/3178876.3186130

[35] Tarleton Gillespie. 2018. *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media.* Yale University Press.

[36] Ajeet Grewal, Jerry Jiang, Gary Lam, Tristan Jung, Lohith Vuddemarri, Quannan Li, Aaditya Landge, and Jimmy Lin. 2018. RecService: Distributed Real-Time Graph Processing at Twitter. In *10th USENIX Workshop on Hot Topics in Cloud Computing (HotCloud 18)*. USENIX Association, Boston, MA. https://www.usenix.org/conference/hotcloud18/presentation/grewal

[37] Alex Hern. 2021. Twitter accidentally blocks users who post the word 'Memphis'. https://www.theguardian.com/technology/2021/mar/15/twitter-accidentally-blocks-users-who-post-the-word-memphis

[38] Michael H. G. Hoffmann. 2020. Reflective Consensus Building on Wicked Problems with the Reflect! Platform. *Science and Engineering Ethics* 26, 2 (2020), 793–819. https://doi.org/10.1007/s11948-019-00132-0

[39] Instagram. 2020. Keeping People Informed, Safe, and Supported on Instagram. *Instagram* (2020). https://about.instagram.com/blog/announcements/coronavirus-keeping-people-safe-informed-and-supported-on-instagram/.

[40] Peter Jachim, Filipo Sharevski, and Paige Treebridge. 2020. TrollHunter [Evader]: Automated Detection [Evasion] of Twitter Trolls During the COVID-19 Pandemic. In *New Security Paradigms Workshop 2020* (Online, USA) *(NSPW '20)*. Association for Computing Machinery, New York, NY, USA, 59–75. https://doi.org/10.1145/3442167.3442169

[41] Jan Jekielek. 2020. Massive, Unexpected Growth on New Free Speech Platform, Bypassing Shadow Bans—Parler CEO John Matze. https://www.theepochtimes.com/massive-unexpected-growth-on-new-free-speech-platform-bypassing-shadow-bans-parler-ceo-john-matze$_2$963526.html

[42] Shagun Jhaver, Darren Scott Appling, Eric Gilbert, and Amy Bruckman. 2019. "Did You Suspect the Post Would Be Removed?": Understanding User Reactions to Content Removals on Reddit. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 192 (Nov. 2019), 33 pages. https://doi.org/10.1145/3359294

[43] Shagun Jhaver, Sucheta Ghoshal, Amy Bruckman, and Eric Gilbert. 2018. Online Harassment and Content Moderation: The Case of Blocklists. *ACM Trans. Comput.-Hum. Interact.* 25, 2, Article 12 (March 2018), 33 pages. https://doi.org/10.1145/3185593

[44] Elena Karahanna, Sean Xin Xu, Yan Xu, and Nan Andy Zhang. 2018. The Needs–Affordances–Features Perspective for the Use of Social Media. *MIS Q.*

42, 3 (Sept. 2018), 737–756. https://doi.org/10.25300/MISQ/2018/11492
[45] Farzaneh Karegar, John Sören Pettersson, and Simone Fischer-Hübner. 2020. The Dilemma of User Engagement in Privacy Notices: Effects of Interaction Modes and Habituation on User Attention. *ACM Trans. Priv. Secur.* 23, 1, Article 5 (Feb. 2020), 38 pages. https://doi.org/10.1145/3372296
[46] Anastasia Kozyreva, Stephan Lewandowsky, and Ralph Hertwig. 2020. Citizens Versus the Internet: Confronting Digital Challaengs With Cognitive Tools. 21 (2020). https://doi.org/10.1177/1529100620946707
[47] Jeanine Krath, Linda Schürmann, and Harald F. O. von Korflesch. 2021. Revealing the theoretical basis of gamification: A systematic review and analysis of theory in research on gamification, serious games and game-based learning. *Computers in Human Behavior* 125 (Dec 2021), 106963. https://doi.org/10.1016/j.chb.2021.106963
[48] Nir Kshetri and Jeffrey Voas. 2017. The Economics of "Fake News". *IT Professional* 19, 6 (2017), 8–12. https://doi.org/10.1109/MITP.2017.4241459
[49] Sang Jib Kwon, Eunil Park, and Ki Joon Kim. 2014. What drives successful social networking services? A comparative analysis of user acceptance of Facebook and Twitter. *The Social Science Journal* 51, 4 (2014), 534–544.
[50] Raph Levien. 1998. Advogado Trust Metric. http://advogato.org/
[51] Nicholas J. Long and Henrietta L. Moore. 01 Mar. 2012. Sociality Revisited: Setting a New Agenda. *The Cambridge Journal of Anthropology* 30, 1 (01 Mar. 2012), 40 – 47.
[52] Ulrik Lyngs, Kai Lukoff, Petr Slovak, William Seymour, Helena Webb, Marina Jirotka, Jun Zhao, Max Van Kleek, and Nigel Shadbolt. 2020. 'I Just Want to Hack Myself to Not Get Distracted': Evaluating Design Interventions for Self-Control on Facebook. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3313831.3376672
[53] Roger C Mayer, James H Davis, and F David Schoorman. 1995. An integrative model of organizational trust. *Academy of management review* 20, 3 (1995), 709–734.
[54] Christian Montag, Bernd Lachmann, Marc Herrlich, and Katharina Zweig. 2019. Addictive Features of Social Media/Messenger Platforms and Freemium Games against the Background of Psychological and Economic Theories. *International Journal of Environmental Research and Public Health* 16, 14 (2019). https://doi.org/10.3390/ijerph16142612
[55] Federico Monti, Fabrizio Frasca, Davide Eynard, Damon Mannion, and Michael M. Bronstein. 2019. Fake News Detection on Social Media using Geometric Deep Learning. arXiv:1902.06673 [cs.SI]
[56] Brendan Nyhan and Jason Reifler. 2010. When corrections fail: The persistence of political misperceptions. *Political Behavior* 32, 2 (2010), 303–330.
[57] Anne Oeldorf-Hirsch, Mike Schmierbach, Alyssa Appelman, and Michael P. Boyle. 2020. The Ineffectiveness of Fact-Checking Labels on News Memes and Articles. *Mass Communication and Society* 23, 5 (2020), 682–704. https://doi.org/10.1080/15205436.2020.1733613 arXiv:https://doi.org/10.1080/15205436.2020.1733613
[58] Sushil K. Oswal. 2019. Breaking the Exclusionary Boundary between User Experience and Access: Steps toward Making UX Inclusive of Users with Disabilities. In *Proceedings of the 37th ACM International Conference on the Design of Communication* (Portland, Oregon) *(SIGDOC '19)*. Association for Computing Machinery, New York, NY, USA, Article 12, 8 pages. https://doi.org/10.1145/3328020.3353957
[59] Marinella Paciello, Carlo Tramontano, Annalaura Nocentini, Roberta Fida, and Ersilia Menesini. 2020. The role of traditional and online moral disengagement on cyberbullying: Do externalising problems make any difference? *Computers in Human Behavior* 103 (2020), 190–198.
[60] Jessica A. Pater, Moon K. Kim, Elizabeth D. Mynatt, and Casey Fiesler. 2016. Characterizations of Online Harassment: Comparing Policies Across Social Media Platforms. In *Proceedings of the 19th International Conference on Supporting Group Work* (Sanibel Island, Florida, USA) *(GROUP '16)*. Association for Computing Machinery, New York, NY, USA, 369–374. https://doi.org/10.1145/2957276.2957297
[61] Emma Peironi, Peter Jachim, Nathaniel Jachim, and Filipo Sharevski. 2021. Parlermonium: A Data-Driven UX Design Evaluation of the Parler Platform. In *Critical Thinking in the Age of Misinformation CHI 2021*.
[62] PEN America. 2021. Defining "Online Abuse": A Glossary of Terms. https://onlineharassmentfieldmanual.pen.org/defining-online-harassment-a-glossary-of-terms/
[63] Gordon Pennycook, Jonathan A. Fugelsang, and Derek J. Koehler. 2015. What makes us think? A three-stage dual-process model of analytic engagement. *Cognitive Psychology* 80 (2015), 34–72. https://doi.org/10.1016/j.cogpsych.2015.05.001
[64] Gordon Pennycook and David G. Rand. 2020. Who falls for fake news? The roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking. *Journal of Personality* 88, 2 (2020), 185–200. https://doi.org/10.1111/jopy.12476 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/jopy.12476
[65] Gordon Pennycook and David G. Rand. 2021. The Psychology of Fake News. *Trends in Cognitive Sciences* 25, 5 (2021), 388–402. https://doi.org/10.1016/j.tics.2021.02.007
[66] Scott R Peppet. 2011. Unraveling privacy: The personal prospectus and the threat of a full-disclosure future. *Nw. UL Rev.* 105 (2011), 1153.
[67] ProSocial Design Inc. 2020. ProSocial Design network. https://www.prosocialdesign.org
[68] Newley Purnell. 2021. Facebook Ends Ban on Posts Asserting Covid-19 Was Man-Made. https://www.wsj.com/articles/facebook-ends-ban-on-posts-asserting-covid-19-was-man-made-11622094890
[69] Bahar Radfar, Karthik Shivaram, and Aron Culotta. 2020. Characterizing Variation in Toxic Language by Social Context. *Proceedings of the International AAAI Conference on Web and Social Media* 14, 1 (May 2020), 959–963. https://ojs.aaai.org/index.php/ICWSM/article/view/7366
[70] Varun Rai and Ariane L. Beck. 2017. Play and learn: Serious games in breaking informational barriers in residential solar energy adoption in the United States. 27 (May 2017), 70–77. https://doi.org/10.1016/j.erss.2017.03.001
[71] Sarah T. Roberts. 2018. Digital detritus: 'Error' and the logic of opacity in social media content moderation. *First Monday* 23, 3 (Mar. 2018). https://doi.org/10.5210/fm.v23i3.8283
[72] Alex Rochefort. 2020. Regulating Social Media Platforms: A Comparative Policy Analysis. *Communication Law and Policy* 25, 2 (2020), 225–260. https://doi.org/10.1080/10811680.2020.1735194 arXiv:https://doi.org/10.1080/10811680.2020.1735194
[73] Jennifer Rose. 2020. To Believe or Not to Believe: an Epistemic Exploration of Fake News, Truth, and the Limits of Knowing. *Postdigital Science and Education* 2, 1 (2020), 202–216. https://doi.org/10.1007/s42438-019-00068-5
[74] Yoel Roth and Nick Pickles. 2020. Updating our approach to misleading information. *Twitter* (2020). https://blog.twitter.com/en$_{u}$s/topics/product/2020/updating-our-approach-to-misleading-information.html.
[75] Joni Salminen, Kathleen Guan, Soon-Gyo Jung, Shammur A. Chowdhury, and Bernard J. Jansen. 2020. *A Literature Review of Quantitative Persona Creation.* Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376502
[76] José Ramón Saura, Daniel Palacios-Marqués, and Agustín Iturricha-Fernández. 2021. Ethical design in social media: Assessing the main performance measurements of user online behavior modification. *Journal of Business Research* 129 (2021), 271–281. https://doi.org/10.1016/j.jbusres.2021.03.001
[77] Filipo Sharevski, Raniem Alsaadi, Peter Jachim, and Emma Pieroni. 2021. Misinformation Warning Labels: Twitter's Soft Moderation Effects on COVID-19 Vaccine Belief Echoes. https://arxiv.org/abs/2104.00779
[78] Filipo Sharevski, Peter Jachim, and Emma Pieroni. 2020. WikipediaBot: Machine Learning Assisted Adversarial Manipulation of Wikipedia Articles. In *Proceedings of DYNAMICS 2020: 2020 Workshop in DYnamic and Novel Advances in Machine Learning and Intelligent Cyber Security* (Lexington, KY). Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3477997.3478008
[79] Qinlan Shen and Carolyn Rose. 2019. The discourse of online content moderation: Investigating polarized user responses to changes in reddit's quarantine policy. In *Proceedings of the Third Workshop on Abusive Language Online*. 58–69.
[80] Qinlan Shen, Michael Yoder, Yohan Jo, and Carolyn Rose. 2018. Perceptions of Censorship and Moderation Bias in Political Debate Forums. *Proceedings of the International AAAI Conference on Web and Social Media* 12, 1 (Jun. 2018). https://ojs.aaai.org/index.php/ICWSM/article/view/15002
[81] Patrick E. Shrout and Joseph L. Rodgers. 2018. Psychology, Science, and Knowledge Construction: Broadening Perspectives from the Replication Crisis. *Annual Review of Psychology* 69, 1 (2018), 487–510.
[82] Jeff Smith. 2017. Designing Against Misinformation. *Medium* (2017). https://medium.com/facebook-design/designing-against-misinformation-e5846b3aa1e2.
[83] Todd Spangler. 2020. Twitter Flags 200 Trump Posts as False or Disputed Since Election Day - Variety. *Variety* (2020). https://variety.com/2020/digital/news/twitter-trump-200-disputed-misleading-claims-election-1234841137/.
[84] Kate Starbird. 2020. How a Crisis Researcher Makes Sense of Covid-19 Misinformation. https://onezero.medium.com/
[85] Richard S. Sutton and Andrew G. Barto. 2018. *Reinforcement Learning, second edition: An Introduction* (second edition ed.). Bradford Books.
[86] Rungting Tu, Peishan Hsieh, and Wenting Feng. 2019. Walking for fun or for "likes"? The impacts of different gamification orientations of fitness apps on consumers' physical activities. *Sport Management Review* 22, 5 (2019), 682–693.
[87] Twitter. 2020. Information Operations. https://transparency.twitter.com/en/reports/information-operations.html
[88] Twitter. 2021. Birdwatch: A community-driven approach to address misinformation on Twitter. https://twitter.github.io/birdwatch/about/overview/
[89] José Van Dijck. 2013. *The culture of connectivity: A critical history of social media.* Oxford University Press.
[90] Teun A Van Dijk. 2009. *Society and discourse: How social contexts influence text and talk.* Cambridge University Press.
[91] Mark Warner, Andreas Gutmann, M. Angela Sasse, and Ann Blandford. 2018. Privacy Unraveling Around Explicit HIV Status Disclosure Fields in the Online Geosocial Hookup App Grindr. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 181 (Nov. 2018), 22 pages. https://doi.org/10.1145/3274450

[92] Brian E. Weeks, Daniel S. Lane, Dam Hee Kim, Slgi S. Lee, and Nojin Kwak. 2017. Incidental Exposure, Selective Exposure, and Political Information Sharing: Integrating Online Exposure Patterns and Expression on Social Media. *Journal of Computer-Mediated Communication* 22, 6 (11 2017), 363–379. https://doi.org/10.1111/jcc4.12199

[93] Sarah Myers West. 2018. Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society* 20, 11 (2018), 4366–4383. https://doi.org/10.1177/1461444818773059 arXiv:https://doi.org/10.1177/1461444818773059

[94] Fiona Westin and Sonia Chiasson. 2019. Opt out of Privacy or "Go Home": Understanding Reluctant Privacy Behaviours through the FoMO-Centric Design Paradigm. In *Proceedings of the New Security Paradigms Workshop* (San Carlos, Costa Rica) *(NSPW '19)*. Association for Computing Machinery, New York, NY, USA, 57–67. https://doi.org/10.1145/3368860.3368865

[95] Felix A Wichmann, Lindsay T Sharpe, and Karl R Gegenfurtner. 2002. The contributions of color to recognition memory for natural scenes. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 28, 3 (2002), 509.

[96] Bernard Williams. 2010. *Truth and truthfulness*. Princeton University Press.

[97] James Williams. 2018. *Stand Out of Our Light: Freedom and Resistance in the Attention Economy*. Cambridge University Press.

[98] Thomas Wood and Ethan Porter. 2019. The elusive backfire effect: Mass attitudes' steadfast factual adherence. *Political Behavior* 41, 1 (2019), 135–163.

[99] Bo Xu, Yun Huang, Haewoon Kwak, and Noshir Contractor. 2013. Structures of Broken Ties: Exploring Unfollow Behavior on Twitter *(CSCW '13)*. Association for Computing Machinery, New York, NY, USA, 871–876. https://doi.org/10.1145/2441776.2441875

[100] Hualong Yang and Dan Li. 2021. Understanding the dark side of gamification health management: A stress perspective. *Information Processing & Management* 58, 5 (Sep 2021), 102649. https://doi.org/10.1016/j.ipm.2021.102649

[101] Zhi Yang, Christo Wilson, Xiao Wang, Tingting Gao, Ben Y. Zhao, and Yafei Dai. 2014. Uncovering Social Network Sybils in the Wild. 8, 1, Article 2 (Feb. 2014), 29 pages. https://doi.org/10.1145/2556609

[102] Waheeb Yaqub, Otari Kakhidze, Morgan L. Brockman, Nasir Memon, and Sameer Patil. 2020. Effects of Credibility Indicators on Social Media News Sharing Intent. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376213

[103] Haifeng Yu, Phillip B. Gibbons, Michael Kaminsky, and Feng Xiao. 2008. SybilLimit: A Near-Optimal Social Network Defense against Sybil Attacks. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*. 3–17. https://doi.org/10.1109/SP.2008.13

[104] Savvas Zannettou, Tristan Caulfield, Emiliano De Cristofaro, Nicolas Kourtelris, Ilias Leontiadis, Michael Sirivianos, Gianluca Stringhini, and Jeremy Blackburn. 2017. The Web Centipede: Understanding How Web Communities Influence Each Other through the Lens of Mainstream and Alternative News Sources. In *Proceedings of the 2017 Internet Measurement Conference* (London, United Kingdom) *(IMC '17)*. Association for Computing Machinery, New York, NY, USA, 405–417. https://doi.org/10.1145/3131365.3131390

[105] Savvas Zannettou, Tristan Caulfield, William Setzer, Michael Sirivianos, Gianluca Stringhini, and Jeremy Blackburn. 2019. Who Let The Trolls Out? Towards Understanding State-Sponsored Trolls. In *Proceedings of the 10th ACM Conference on Web Science* (Boston, Massachusetts, USA) *(WebSci '19)*. Association for Computing Machinery, New York, NY, USA, 353–362. https://doi.org/10.1145/3292522.3326016

[106] Savvas Zannettou, Michael Sirivianos, Jeremy Blackburn, and Nicolas Kourtellis. 2019. The Web of False Information: Rumors, Fake News, Hoaxes, Clickbait, and Various Other Shenanigans. *J. Data and Information Quality* 11, 3, Article 10 (May 2019), 37 pages. https://doi.org/10.1145/3309699

[107] Daniel Yue Zhang, Rungang Han, Dong Wang, and Chao Huang. 2016. On robust truth discovery in sparse social media sensing. In *2016 IEEE International Conference on Big Data (Big Data)*. 1076–1081. https://doi.org/10.1109/BigData.2016.7840710

APPENDIX

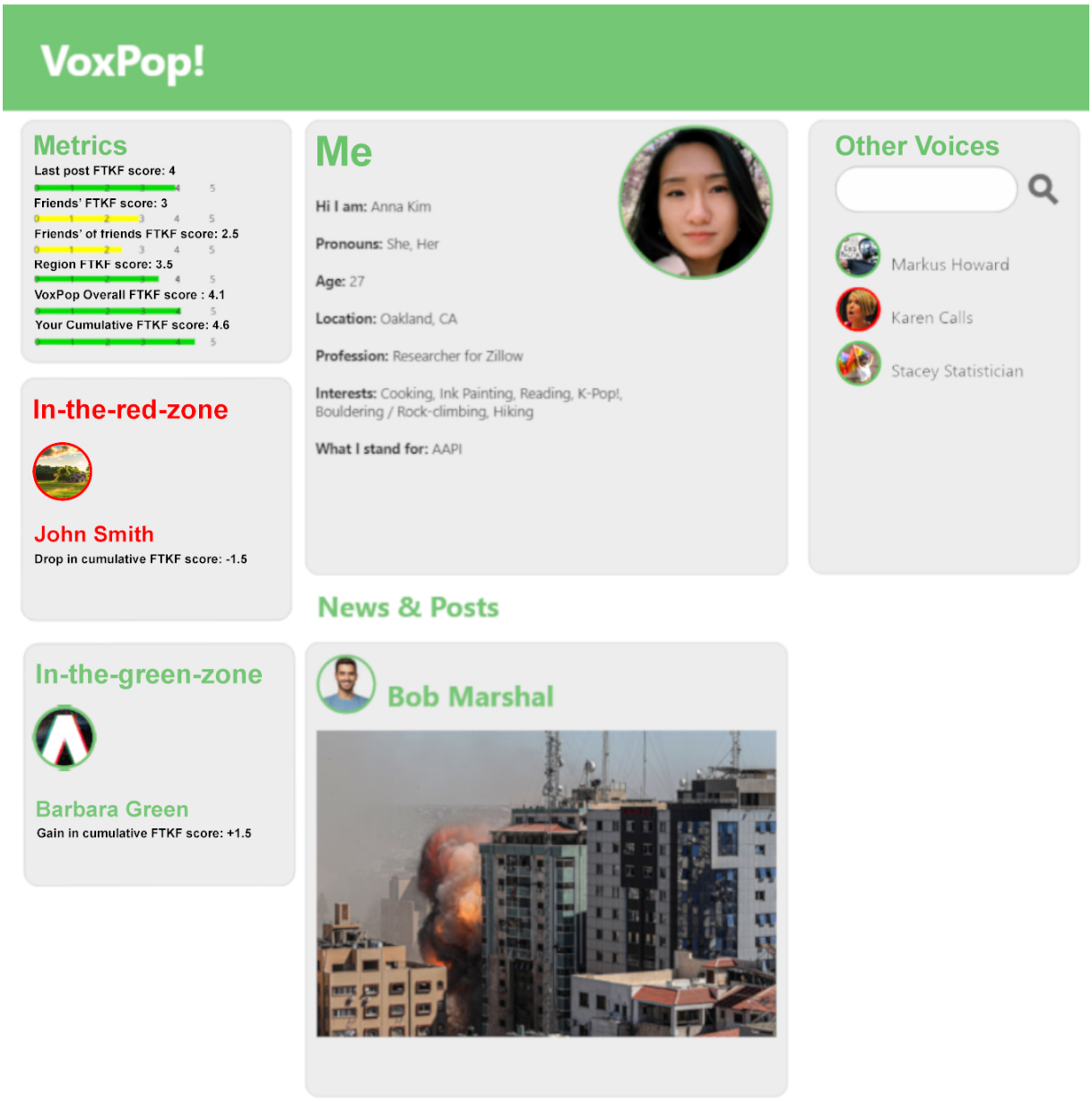

Figure 11: An example homepage in which the user's C-FTFK score is high, therefore the banner appears in green

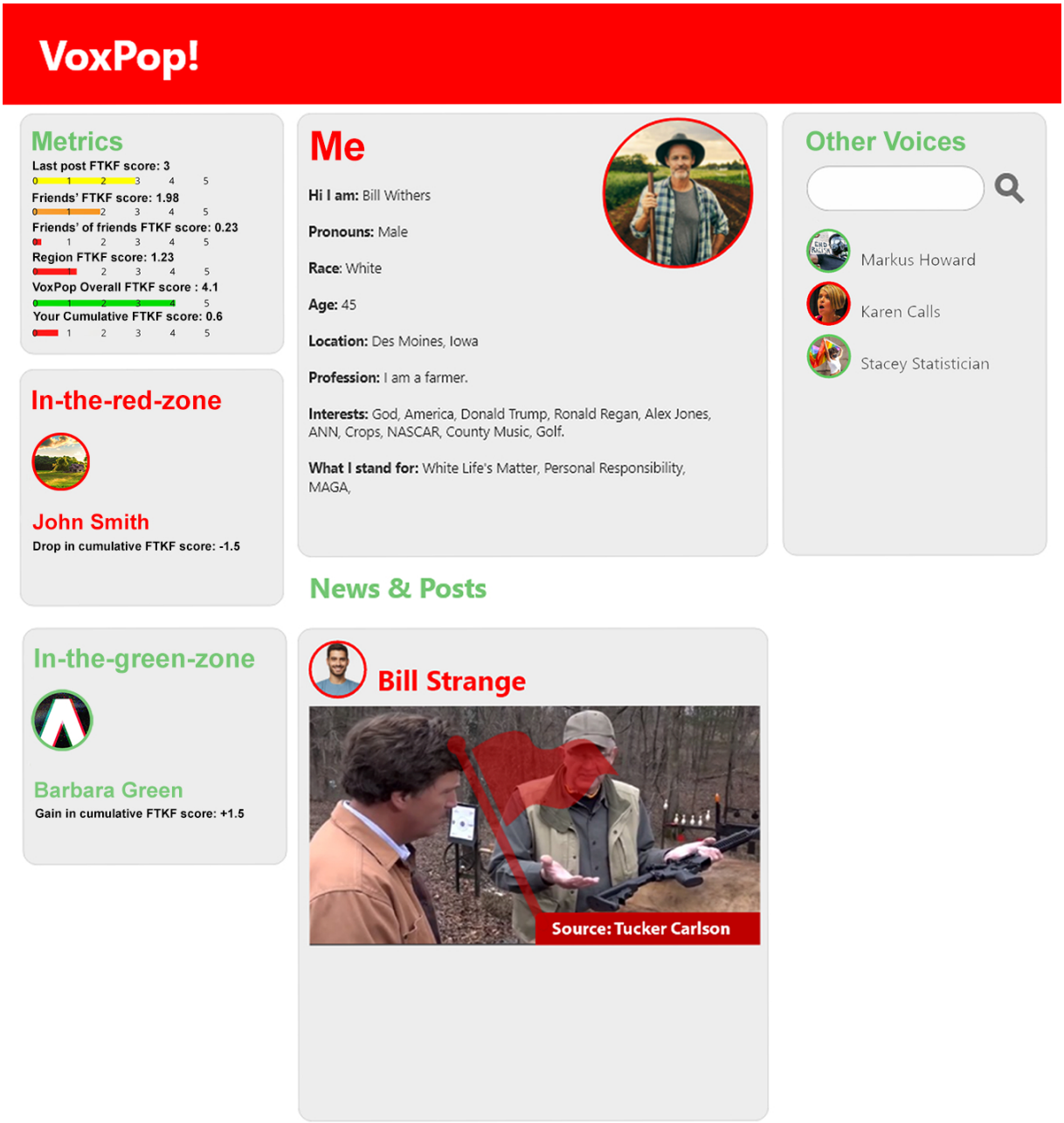

Figure 12: An example homepage in which the user's C-FTFK score is low, therefore the banner appears in red