Who watches the Birdwatchers? Sociotechnical vulnerabilities in Twitter's content contextualisation

Garfield Benjamin[1]

Solent University, Southampton, UK
garfield.benjamin@solent.ac.uk

Abstract. At the start of 2021, Twitter launched a closed US pilot of Birdwatch, a new system for content moderation that seeks to aid in promoting credible information online by giving users the opportunity to add context to misleading tweets. The pilot shows awareness of the importance of context, and the challenges the system will face. But the mitigations against these vulnerabilities of Birdwatch can exacerbate wider societal vulnerabilities created by Birdwatch. This article examines how Twitter presents the Birdwatch system, outlines a taxonomy of potential sociotechnical vulnerabilities, and situates these risks within broader social issues. We highlight the importance of watching the watchers, not only in terms of those using and potentially manipulating Birdwatch, but also the way Twitter is developing the system and their wider decision-making processes that impact on public discourse.

Keywords: Twitter · Sociotechnical · Vulnerabilities · Online content.

1 Background

Against the backdrop of ongoing problems with social media platforms and online content moderation, Twitter's Birdwatch system enables users to add 'notes' to tweets, in order to add context, fact-check or other explanation aimed at promoting more credible information online. At face value, the initiative seems positive, particularly following waves of not only political but also public health issues caused by misleading information circulating through social media, and the flurry of forthcoming regulatory changes in response [19]. If anything, the main criticism we could level is that it should have been introduced sooner.

It is reassuring - maybe even a little surprising - to see a social media platform introducing important concepts like context. Researchers have been emphasising the importance of context in issues of data, privacy and algorithms [16,50,7,23,48,51,44,22]. Seeing a major platform embrace the nuances of context that surround all information is an important step, particularly if it involves users being able to contribute to that context.

But there are clear potential problems with Birdwatch: in its conceptualisation, in the design of the pilot, and in the planned path towards global rollout.

Any system engaging with complex sociotechnical issues will likely be riddled with bugs and vulnerabilities. Sociotechnical problems generally have no single concrete 'fix'. Controlling and moderating content online has competing arguments and interests that resist a final resolution to these issues.

Grimmelmann calls content moderation "the governance mechanisms that structure participation in a community to facilitate cooperation and prevent abuse" [31], including processes of designing and moderating. But moderation is a fiendishly difficult problem to tackle, riddled with issues of power and labour [54], and often involving competing interests, ideals and methods.

Content moderation, and the inadequate approaches by platforms thus far, have come under much scrutiny and criticism [54]. This includes issues of ghost work by exploited and undervalued moderators [30,46], hidden decision-making by corporate interests [26], and defining what constitutes the public sphere [66]. In practical terms, content moderation has been criticised for negative or contradictory user perceptions [49,9], inspiring practices of avoiding moderation techniques [25], and promoting further engagement (even if negative) with harmful or false content [67]. Content moderation demonstrates difficulties in regulation, in terms of defining harms and the hidden algorithmic back-end [13], the informal interactions between platforms [18] or with states and NGOs [27], and the need to support international human rights across legislative settings [41].

Birdwatch adds in further issues by inserting community-based mechanisms [40] within the organised processes of "commercial content moderation" [54] that form one of the key commodities provided by online platforms [26]. But how far is Birdwatch a community-based approach? It is certainly crowdsourced, and makes impressive claims for transparency. But it risks being at best a partial offloading of labour and responsibility to the community while the platform itself continues to benefit from providing user-generated content.

Given the "systemic vulnerabilities in our democracy" [35] created by social media platforms, we need to keep a careful watch over Birdwatch. We need to ask intersectional questions of "who" [16]: who benefits from the platform? Whose voice will be (further) amplified? Who will be (further) marginalised? Who is exploited to generate the data for the platform to function? Whose interests and biases are designing the system? We must also keep watch for other motives at work. Is Birdwatch's contextualisation even content moderation? Or is it another step towards promoting platform engagement to distract from outside regulation, leaving the underlying power of Twitter unchecked?

In her critique of racialised histories of surveillance, Browne [10] emphasises the need to watch the watchers: sousveillance, the surveillance *of* authorities rather than *by* them, is both a critique and a practice. Power comes not only from watching authority but by using insights into surveillance to find ways of resisting. With the fraught intertwining of platform, content and users, it is imperative that we build on these practices to watch the Birdwatchers. This involves not only users watching other users but challenging the platform and dominant voices who control the metricising narratives of online media.

This paper examines how Twitter is constructing and presenting the Birdwatch trials, including the stated values, challenges and mitigations. A taxonomy of potential sociotechnical attacks is outlined, including not only vulnerabilities of the Birdwatch system but also societal vulnerabilities created by Birdwatch. These vulnerabilities are situated in a discussion of broader societal impacts. The article concludes by questioning what role Birdwatch is trying to play, for society and for Twitter. It emphasises the need to not only watch the Birdwatchers, but to scrutinise platform priorities in controlling content and discourse online.

2 Birdwatch

Birdwatch operates as follows [61]:

- A user posts a tweet containing claims that may be false or taken out of context;
- Other users add 'notes' to these tweets to provide context or evidence to explain or counter the claims made in the tweet, including a reason selected from a list of options;
- Further users upvote or downvote notes, selecting a reason from a list of options;
- Notes are ranked according to votes, influencing their visibility in relation to the tweet.

The aim is that users can then assess the credibility of a source themselves, based in crowd-sourced and crowd-rated contextualisation.

Notes are very different from replies. Notes are not a thread following a tweet, but exist alongside, aimed at providing context in a more authoritative way. Compared to anti-abuse mechanisms like hiding or restricting replies, notes are less about direct communication between the tweeter and note writer. They are between a note writer and the audience, and so intervene between the tweet and its reception.

As already noted, there is a lot to be positive about here: emphasising context; emphasising evidence; shifting discussion to around tweets rather than in antagonistic or abusive replies; collective assessment of notes; differentiating reasons for a note and how a note is ranked. There is also an open channel for feedback to Birdwatch, including to highlight risks and express concerns [61], and they show a desire to respond and engage, at least to some degree [1]. This echoes positive moves in other departments, such as the responsible machine learning team [63]. But each of the positives of Birdwatch also brings risks of abuse and wider societal impact, and the design of the system raises further questions we must examine before it becomes normalised.

Visibility is an essential point to consider. Which tweets will register for potential notes? How are notes made visible? These are important questions. Tweets will need "100 total likes plus retweets" [61] in order to appear in the Birdwatch system. This opens up some avenues of avoiding scrutiny either through obscurity, or through more directed means. Iterating tweets across bots

to lower like/retweet counts could reduce the likelihood of it triggering Birdwatch (or simply making the same misleading information more widely seen). Followers pasting (perhaps with minor changes) the same information rather than retweeting could achieve similar results. Images of text could also be used for this purpose. However, these attacks would not scale well to those with large numbers of followers or the hopes of "viral" attention, and would require a greater deal of coordination and participation.

The note ranking system has mechanisms to prevent abuse, but these often shift attacks to other forms. Notes initially register as "Needs more ratings" until 5 or more votes have been received, at which point it may be shifted in the ranking system and have "Currently rated helpful" or "Currently rated not helpful" labels added [61]. During the pilot, rankings are calculated at regular intervals rather than changing in real-time. This might prevent a misleading or malicious note jumping to the top of the feed, but it also introduces what amount to timing attacks. Once a note is added, followers, sockpuppet or bot accounts could upvote it, and then it will (for a certain time at least) be mislabelled as helpful. Even if these things can be rectified by further votes, it relies on numerous assumptions: that the community can and should be willing to take on this role; that further follower/bot coordination isn't going to counter legitimate votes; that even a short time attached to a viral tweet isn't enough to sow misinformation or abuse, particularly if it crosses to a different platform.

Notes and note votes have lists of reasons attached. This to some extent adds further context at a glance. But it also adds another form of context attack. By targeting less used (or even irrelevant) reasons, a note can be made to seem like a different opinion or "diverse perspective". The ranking system is designed to seek out such diversity - different reasons for context, and notes coming from users who have engaged with different accounts or tweets. But it could also promote misleading notes that game the categorisation system.

This highlights an underlying problem with algorithmically sorting notes, and demonstrates how Birdwatch is still trying to turn context into metrics in order to deal with the issues of content moderation at scale. The note ranking algorithm is less transparent, and will need particular scrutiny as the pilot and system develop.

The ability to sign up for the current pilot is restricted to the US [61]. This introduces the potential for data pollution attacks. If these are intentional, it feeds into broader concerns of disproportionate influence by certain actors (perhaps mobilising bots or large numbers of followers). But it also leads to unintentional data poisoning in Birdwatch's design. If the system is shaped too tightly by US linguistic, social, cultural, political and economic concerns, a context attack develops by imposing these norms on other geographical locations when the system is rolled out globally.

The US context is also applicable to the ability to download Birdwatch data for research and accountability purposes [61]. Twitter has long been considered a "model organism" [58] for its publicly available data for research. But given the intention to deploy the system globally, the US-only access excludes wider per-

spectives from the development of the system. The data itself risks adding only faux transparency, as the ranking algorithms themselves remain more opaque, as do the ultimate decision-making processes within Twitter. It seems they are seeking external validation and fixes of vulnerabilities within Birdwatch rather than full sociotechnical audit of the impact of the system itself.

Within the US, there are limited places on the pilot, prioritising again those "diverse voices" by selecting accounts that interact with different tweets or other users on Twitter [61]. This is an attempt to stop bot-based attacks and promote more diverse data. But to gain access, you also need to verify your phone and email, and not have any recent notice of Twitter rules violations.

While these measures may prevent bot and sockpuppet attacks, in mitigating these vulnerabilities they introduce exclusion of those who may need to remain anonymous (abuse victims or whistleblowers, for example), or those who don't have access to the levels of verification needed. The need to sign up with a trusted US-based phone carrier and enact 2-factor authentication may exclude those with less access to technology based on cost or sharing devices. It will therefore likely prioritise already dominant voices and those who have the privilege of being able to use their real names on Twitter without fear of abuse. This shapes the context with an escalation attack in which loud voices get even louder. Mitigating this will rely on the success of the diverse engagement measures.

How is Twitter designing these measures? What would register as success of the system? The company outlines the Birdwatch values as "contribute to build understanding", "act in good faith" and "be helpful, even to those who disagree" [61]. At first glance, these seem reasonable. But, firstly, these are largely expectations placed on users rather than expressing the values of Birdwatch as an initiative. And, secondly, there are holes.

Language use can be easily gamed. Abusive comments could be concealed beneath a veneer of clear or scientific authority, and even citing evidence does not certify the validity of the evidence cited. Bots could be easily trained to write just within the guidelines while still being varied enough to count as different perspectives. Similarly, "baiting" other users into using language that triggers automated toxicity filters has already been employed on gaming platforms [53], and these techniques could be used to game notes.

There are also assumptions and potential data bias about such language. Will the automated parts of the system continue to misunderstand the context? Would critically discussing hateful language lead to a note being itself considered hateful, even if the abusive tweet it discusses didn't? The data used to define these triggers will be important [6,55,17], particularly when it comes to context shifts.

There are fundamental tensions in Birdwatch managing contextualisation through a metricised system. And relying on good faith on social media is almost certainly naive when it comes to a global platform with stakes in preventing abuse and promoting public discourse.

Twitter is aware of many of these challenges [61], and some of the measures already outlined are there to try to mitigate potential attacks. They state that

Birdwatch will add to rather than replacing existing misinformation measures. The aim of the pilot process is to experiment with "new ways to address adversarial coordination" to prevent "coordinated mass attacks" [61]. While these may work in the sandboxed pilot stage, it remains to be seen whether any solutions scale to global rollout across different contexts.

There are issues with Twitter's responses to the acknowledged challenges. Though they may mitigate vulnerabilities within the Birdwatch system, in doing so they may also contribute to further societal risks caused by Birdwatch. Diverse engagement systems offer some potential, but metricising diversity will always be problematic.

Similarly, the parallel reputation system for helpful Birdwatchers (separate from number of followers on Twitter itself, for example), not only risks adding further burden to marginalised communities but simply shifts the gaming. It may create more work to initiate an effective attack, but it is not much beyond trivial for major actors - particularly authoritarian regimes outside the US-EU zone that (so far) receive much less attention from social media platforms. Again, context is key, and the work of critical journalists, researchers and rights advocates paying close attention to global contexts [64,3,21,62,65] (to name just a few) will be key.

The sandboxed trial itself raises methodological questions. The aim of preventing bots from sabotaging the trial is admirable, but any wider implementation will inevitably have to deal with bots. Including bots in the trial would provide a far more realistic view of what challenges will arise. Twitter's choice to test in this way relies on a 'best-case' context of sorts. This sidesteps many of the potential problems in the realities of online platforms, and of course limits the narratives to the already dominant US context. An alternative approach would be an open global trial, but maintaining the Birdwatch platform sandbox so that notes and rankings do not appear on tweets until thorough testing, analysis and consultation have been conducted.

Twitter acknowledges the difficulties in global contexts for expanding Birdwatch, including cultural differences and the specificities of misleading information in different settings - and they have stated that they will take into account these considerations as they expand [1]. But the questions remain of who decides? When will the pilot be expanded, and what measures of success are required first? Where will it be expanded to - other Anglophone or North America-European settings? When will the pilot become local or global implementation, and will the same attention be paid to continuing development?

There is also the fundamental question of would they decide to stop? Twitter's responsible machine learning team shows further reassurance here, giving users the option to turn off image-cropping algorithms that have been proven to be biased [63]. But would they make the leap to abandoning Birdwatch if it fails? And who would it need to fail for in order for it to be stopped? This is likely more a matter of who is failed by Birdwatch, and it will likely be those that platforms such as Twitter have already failed.

3 Sociotechnical vulnerabilities

Algorithmic content moderation is about people [54], and any discussion of its effects must be embedded in social as well as technical concerns. For example, [28] identify how algorithmic content moderation faces a series of not only technical but political challenges. We argue to go even further and not separate these two spaces.

The technical is political; the political is technical. This is shown in algorithms that discriminate based on race and/or gender [11,38,32] and the proposal that categories of marginalisation such as race are themselves constructed as technologies [8]. We argue for talking in interwoven sociotechnical terms, particularly when assessing vulnerabilities in the integrity of public discourse and control over social narratives.

This approach differs from, for example, social engineering in that it is not concerned with the use of human or social side-channels as alternative vectors to attack technical systems. And it goes beyond looking at the social effects of technological vulnerabilities. Instead, it focuses on the use of both social and technical channels in order to attack systems that are themselves social. It employs the languages of vulnerability analysis and critical theories in order to examine potential social and societal harms, whether that is abuse of individuals, marginalisation of specific groups, or risks to public discourse and democracy. Sociotechnical vulnerabilities are about the distribution of power, equity and justice.

Applying these social vulnerabilities to Birdwatch generates two broad categories. Firstly, there are vulnerabilities of the Birdwatch system. This includes all the ways that Birdwatch can be manipulated or abused, from avoiding the system to weaponising notes to gaming rankings. Secondly, there are vulnerabilities by Birdwatch. This includes the risks to public discourse and other areas of societal concern created by the use of the system, even as intended.

Figure 1 shows how these types of vulnerability interact, highlighting how the mitigations of the vulnerabilities of Birdwatch can exacerbate the vulnerabilities created by Birdwatch. It is important, therefore, to carefully review how Twitter acknowledges and mitigates the challenges of developing the Birdwatch system. This is not just examining how effective they might be, or further loopholes for attack, but how the mitigations themselves feed back into how Birdwatch creates vulnerabilities for society.

An initial response to these broad categories of vulnerabilities should include a requirement for continual external audit. This should be enforced by regulators and could be conducted using methods following, for example, the Platform Governance Observatory (Suzor, 2020). This process highlights the importance of wider access to the data downloads and the underlying algorithms. The stakes of the system being rolled out globally suggest a narrowing of access to within the US - matching the terms of the pilot - could have long term harmful effects or necessitate those with access ‘leaking’ it to researchers, activist and rights advocacy groups, and regulators, in other geographical contexts.

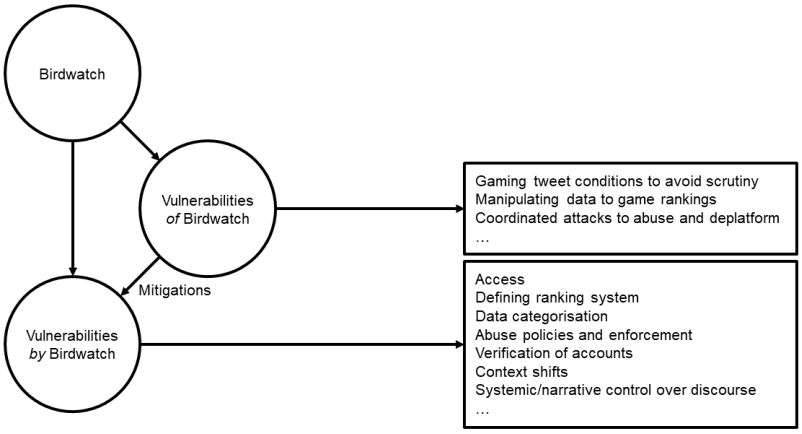

Fig. 1. Vulnerability structure of Birdwatch.

3.1 Taxonomy of sociotechnical vulnerabilities

Many of the vulnerabilities have been outlined in the analysis of Birdwatch above, but it is useful to consider how different vulnerabilities are related. Figure 2 shows a proposed taxonomy of the social vulnerabilities of and by Birdwatch. These are traced from the target area, through the various modes of attack, to the societal harms the attack creates or worsens. The taxonomy embodies the sociotechnical approach of this article, combining elements of, for example, the NIST adversarial machine learning taxonomy [57] as well as Citron and Solove's typology of privacy harms [12]. The taxonomy emerges from the reading of Birdwatch presented above, systematically analysing the parts of the system and Twitter's approach to developing and presenting it. The full description of these elements can be found in Appendix A.

3.2 Ranking vulnerabilities

In technical settings, it is useful to rank attacks according to considerations such as likelihood or severity, particularly when taxonomising. However, in socially-embedded contexts such as online content moderation, this can prove impossible without introducing further assumptions and exclusions. Jiang et al. [37] show that "Platform abuse and spam" ranked consistently low in user perceptions of severity of online harms, and even "Mass scale harms" such as terrorism ranked lower than specific cases of harmful content, although these mass scale harms remained present in participants' concerns. The study highlights the need to move outside of US-centric perspectives when moderating content, particularly when making assumptions about its severity. Any collapsing of complex social

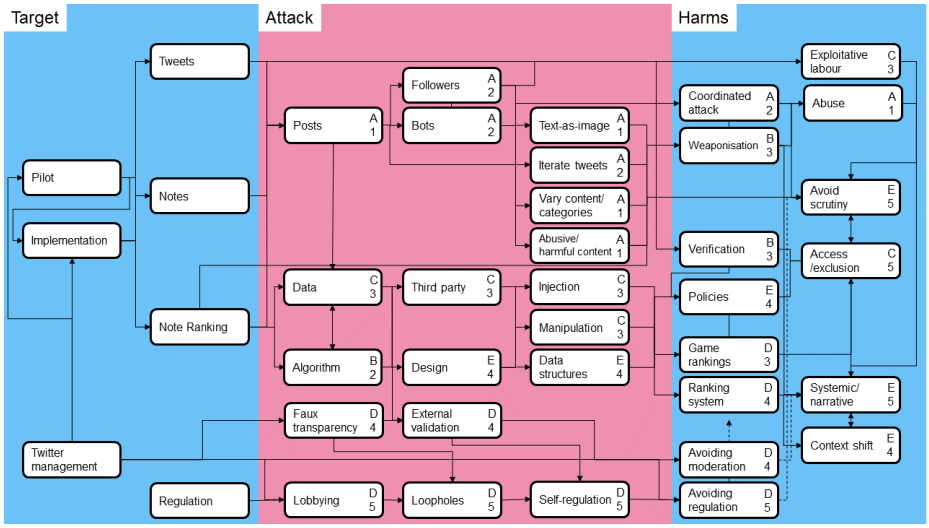

Fig. 2. Taxonomy of attacks.

contexts into restrictive metrics will erase important details, and runs counter to the aim of examining the implications of online platforms. However, there are useful indicators that can assist in the examination of different types of attacks and harms, particularly against individuals. Indeed, this tension is a dilemma within Birdwatch itself: mitigating individual harms while exacerbating systemic harms does little to resolve the narratives in which the individual harms are enabled. To this end, rather than the perceptual and evasive labels of severity or likelihood (which in any case are so unevenly distributed as to be of little use without further qualification), we instead rate attack and harm elements within the taxonomy according to the scale and timeframe of their effects. This is assigned based on the primary impact identified in the analysis above, and presents us with two values for each attack:

Scale

- A Specific individual
- B Specific types of individual (e.g. based on protected characteristics)
- C Specific community
- D Public discourse in a specific context (e.g. field or geographical area)
- E Systemic/principles (e.g. democracy, discrimination, censorship)

Timeframe

1. Immediate (e.g. on posting)
2. Short term (e.g. sharing, including viral propagation before removal)
3. Mid term (e.g. posts over time)
4. Long term (e.g. impacting design iterations)
5. Persistent (e.g. ongoing narratives, systemic decision-making).

This allows us to further identify which vulnerabilities can be mitigated within Birdwatch, and which require external intervention (such as the potential systemic vulnerabilities created by the narratives that Birdwatch enables).

3.3 Limitations

Inevitably a taxonomy of this kind will not be exhaustive. Emergent practices and evolving system design will create and mitigate new vulnerabilities over time, and vulnerabilities will vary in effect in different contexts. The aim is to provide an approach for discussing how different types of vulnerability can be combined to cause social and societal harms. The limitations to the taxonomy presented here echo the limitations of Birdwatch itself. They are likely to be useful within specific contexts (with certain types of information such as medical advice, for example) rather than as a blanket tool across the entirety of the Twitter platform. The examples discussed below identify unequal harms and limitations that apply in specific contexts or to specific users. This further limits the ability to assign concrete ratings for harms beyond speculative potential or assigning values based on either averages or best-/worst-case scenarios.

4 Examples

We now examine one of the processes in which these social attacks might be mobilised. Consider the example of a scientist, who happens to identify as a Black woman, posting on Twitter about some new research findings or commenting on an issue of public concern well within her area of expertise. The example is familiar from throughout the Covid-19 pandemic in particular, but also applies more widely when minoritised groups (whether women, racialised people, trans or queer people, or disabled people, just to name a few) share their expertise and/or lived experience in fields from computer science to biology to politics to philosophy. Figure 3 shows a possible process of attack, which aligns with the outline of "The Abuse and Misogynoir Playbook" [60].

- Firstly, a *contribution* is made. This is the victim posting on Twitter.
- This is met with *disbelief* by the abuser (which may be someone already in a position of power and/or platform, including senior academics, public officials/candidates, celebrities or other public figures) and their followers, leading the abuser to write a bad faith note containing falsehoods. The note itself, while based in opinion and falsehood, will likely be crafted to appear to follow the Birdwatch guidelines: no direct insult; appealing to a sense of commonly held knowledge or disputed/beside-the-point technicalities/"whataboutism".
- This is followed by a combination of *dismissal*, *gaslighting* and *discrediting*. These could be achieved by the abuser's followers piling on with further false or bad faith notes (many of which may get flagged as abusive), followed by the abuser's followers and/or bots upvoting the false notes while downvoting any notes written in support of the victim.
- The result of this is the *erasure* of the victim through devaluing the original contribution and hijacking the narrative. Even if others support the original post with supportive notes, with enough negative engagement and conflicting false narratives, it will likely end up being downranked by the Twitter algorithm.

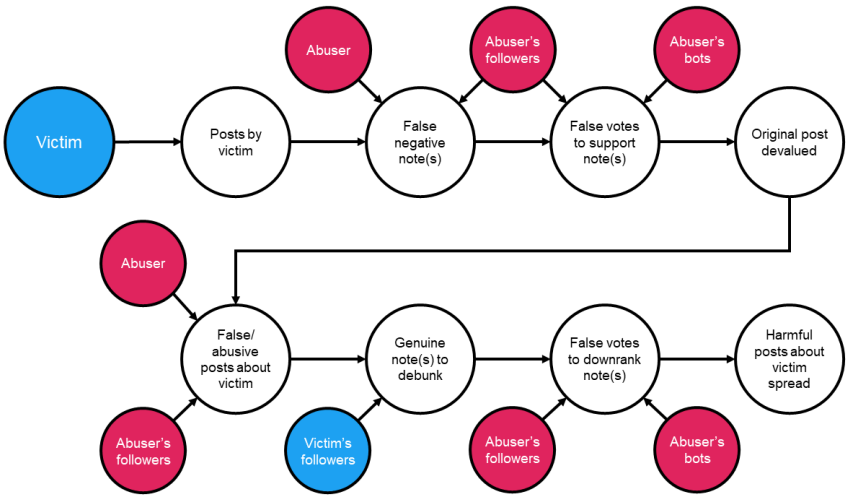

Fig. 3. Combining data poisoning, gaming the algorithm and coordinated attacks.

– This is furthered into the final misogynoir move, *revisionism*. Here the whole process happens in the negative: the abuser and their followers write false or abusive posts about the victim or their contribution; genuine notes by the victim's followers or others in the community are further gaslit and discredited by false downvotes by the abuser's followers, while continuing to upvote or add positive notes about the abusive content; and finally the harmful and abusive posts can be either spread further than the original contribution or entered into enough of a contested narrative that the lines between contribution and abuse become sufficiently blurred to devalue both.

Either way, data poisoning (false notes, false votes), gaming the algorithm (affecting the rankings), and coordinated attacks (whether by followers, bots or a combination) have led to abuse, weaponisation, exclusion and a perpetuation of oppressive narratives.

In this example, a number of different types of agent are mobilised. This includes: a human abuser, likely a public figure or other account with large numbers of followers; the human followers of that account, each with their own motives and patterns of behaviour; bots controlled by the abuser, their followers, or third parties seeking to capitalise on abuse, again for a variety of motives. There are also other different interactions and roles between agents to take into account. Figure 4 shows how different types of Twitter account could take on different roles as part of a coordinated mass attack.

This includes a range of bots and bot-human hybrids, building on Gorwa and Guilbeault's typology [29]. For example, sockpuppet accounts - additional and/or anonymous accounts often used specifically for trolling and abuse - could be used by abusers to try to avoid responsibility. Or, followers and bots could

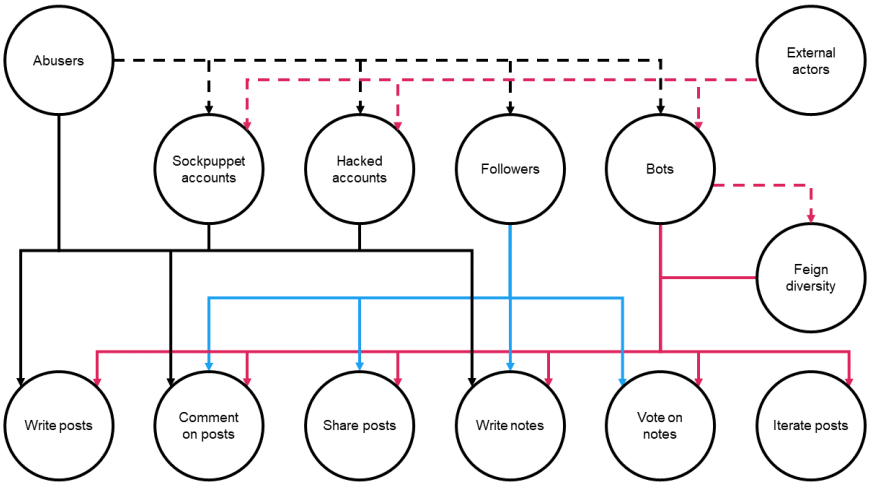

Fig. 4. Actors.

take on crowdturfing roles; the (often paid) fake reviewers that appear on online shopping or trust/review sites could be repurposed to write fake notes. Many existing forms of bots and collective human action could be mobilised in these different roles.

These roles could also manipulate Twitter's existing mitigations of the risks of Birdwatch. An automated or crowd-sourced trolling army could easily be created to spread across different engagement or community types. By establishing different patterns of interaction with different communities and types of content, a range of accounts could be created that the system would identify as coming from "diverse voices". Even a dedicated trolling sockpuppet that interacts solely with people from a specific marginalised group could appear to the system as belonging to or representing that group.

Particularly at scale, Twitter and Birdwatch's automated systems will not be able to tell genuine engagement as part of a community from a targeted harasser. Hijacked accounts can also be used for this purpose. These could then be brought into action as a supposedly representative set of opinions all saying the same thing for malicious and/or abusive ends.

Following existing social media information warfare techniques [39], it is easy to imagine a paid-for service using these different attacks on Birdwatch. Harassment and misinformation as a service is already a problem, and can be particularly damaging to political candidates, especially those from marginalised groups. The diverse voices Birdwatch aims to emphasise could be overridden in a hegemonising attack that collapses the aims of contextualisation.

The ability of different agents to target different aspects of the widening range of interactions - notes and votes as well as tweets - impacts on pathways

of abuse. Birdwatch shifts not only where abuse can occur but also the responses to abuse. Because notes are intended for audiences of tweets, rather than being directed at the original tweeter as is the case with replies, they challenge existing methods of mitigation and recourse.

For example, note-writers may be able to operate outside of the blocking mechanisms that can be used to prevent malicious or abusive replies. This also reduces the effectiveness of tools like automated collective blocklists [24]. Removing tools developed by marginalised communities contributes to the harm of placing extra burden on those communities to rectify content moderation issues.

The wider sociotechnical vulnerabilities are evident in the risks of context shift. Despite being focused on contextualising individual tweets, Birdwatch runs the risk of imposing certain media and power structures onto other contexts. Figure 5 displays some of the aspects of this context shift vulnerability. Structural and cultural contexts often do not translate, and a major risk of Birdwatch is that the US pilot will entrench specific values and priorities. This has implications for imposing and erasing certain norms of public discourse on social media, and for removing agency from global communities.

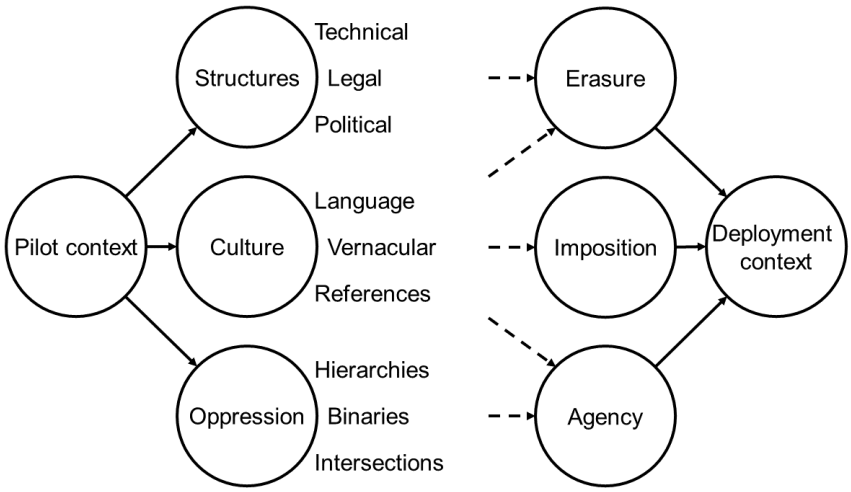

Fig. 5. Context shift.

We need to ask: whose context is Birdwatch adding? Without meaningful engagement with other communities during the development phases, it is likely to perpetuate the dominance of US-centric platforms. Public discourse in other geographical or international contexts is made vulnerable to further colonial (and other forms of) domination.

5 Discussion

A difference between conventional technical vulnerabilities - "hacking" - and social vulnerabilities is that there is often no state of being "fixed" for social systems like content moderation. This is true both over time - contexts, norms and needs may change as society and platforms develop - and in terms of solutionism - there is no final finished state of content moderation, for example, it will always be an ongoing process. These points are somewhat true of technical vulnerabilities - cyber security is always to some extent a matter of constant "patching" or being "secure enough" for the given needs and context - but a central tenet of finding solutions for social vulnerabilities is a resistance to seeing them as closable problems.

The reading of Birdwatch and the subsequent taxonomy suggests that closing social vulnerabilities of Birdwatch tends to create social vulnerabilities by Birdwatch. This is particularly the case with mitigations such as verification, which introduces issues of identification (leading to potential targeting and further abuse) previously not present on Twitter.

Citron and Solove [12] emphasise the differences between the goals of punishing harms - often compensation, deterrence and equity - and the remedies to the harms themselves. This is further evident here where Birdwatch sanctions of specific accounts in order to protect the credibility of the system lead to increased systemic vulnerabilities in public discourse and democratic processes increasingly controlled by Twitter's narrative.

These concerns highlight the importance of watching the watchers, not just those on Birdwatch but the decision-making and regulatory process behind it. And watching itself needs to be one step towards action. Watching does not prevent single incidents, but when it comes to systemic issues like social media and public discourse, it can lead to recourse for specific victims, sanctions for abusers, and bring evidence to light that can help redefine the system as it develops.

There are other aspects of content moderation to consider with Birdwatch. Does it even count as moderation? The more deliberative framing of contextualisation is perhaps a separate process. It is not aimed at fixing specific pieces of content, but at supporting better discourse more generally. It moves beyond what Common [15] describes as an "efficiency narrative" in the enforcement of content moderation. Indeed, moving away from the concerns raised by [12], Birdwatch is not even really about enforcement. But does this different function detract from the need to address issues with specific pieces of content?

We could see Birdwatch as Twitter taking a longer, more constructive approach. The role of contextualisation could contribute to broader aims of improving public discourse on social media. It allows space for restorative justice approaches to moderation [34], as well as opportunities for other techniques such as digital juries [20]. Perhaps Birdwatch is a first step away from enforcement as a narrative, and a rejection of the idea that there is a quick fix for online content and online harms?

But what of the more immediate harms that will continue in the meantime? Responding to specific harmful tweets will still rely on tactics such as blocklists which have been shown to be problematic and inadequate [36]. Meanwhile, journalists and rights advocates may still have the visually shocking content they rely on blocked outright by content moderation before context can be added [4]. However, representation and narrative are important.

It may be that such posts should be blocked to stop perpetuating victim or deficit narratives, like the calls to stop sharing footage of Black men in the US being shot by police as it performs and perpetuates the expectation and normalisation of such acts. Contextualisation, moderation and refusal each have their place in making discourse on social media not just more credible but also more equitable. Power must be taken into account, not just authority.

This leads us back to key questions over what role contextualisation and community are playing on Birdwatch. The mechanisms seem more effective on fact-based issues, like their example of "whales are not actually mammals" [61], which can be readily countered with scientific context according to Birdwatch's guidelines. But the major risks and vulnerabilities have greater impact when it is social issues at stake. Hate speech and misrepresentation of gender (including trans), race, disability and other forms of discrimination require greater context but are more vulnerable to oppressive dominant narratives and attacks.

There are significant issues with relying on public fora as a marketplace of ideas [43], leading to further concerns with the more general gamification of social media discourse [45] and the individualising effects this is likely to have on Birdwatch [42]. Reducing context back to metrics could alienate the people and communities doing the work. Birdwatch claims to be community-based, but is this true?

Twitter introduced Birdwatch as "community-based" [14], citing research on crowd-sourcing to tackle misinformation [52,47,2]. It is worth noting the overlap in authors of two of these papers, and the fact that [2] has not yet been peer-reviewed. But these papers present a significantly narrower scope than the aim of creating credible public discourse.

While it may to some extent avoid efficiency narratives, Birdwatch still feeds into the related and equally problematic narratives of scale [33]. Neil Turkewitz and Emily Bell highlight how "scale doesn't scale", or at least that "harms scale, solutions not so much" [59]. If "Journalism with high civic value [...] is discriminated against by a system that favors scale and shareability" [5], is credible information simply incompatible with Twitter as a medium? Or does Birdwatch signal a shift in this narrative? It at least attempts to reach the combination of human and algorithmic solutions that Bell and others propose, but the combination needs interrogating. Is it trying to humanise the visibility algorithms? Or does it end up algorithmising human users?

Birdwatch may be crowd-sourced, but that is not the same thing as being community-based. Where is the agency for marginalised individuals or communities? Decision-making is still controlled centrally by Twitter, a private platform. It seems more like they are getting extra fact-checking done for free by users,

rather than providing the service of moderation (which, as [26] asserts, is a key function or commodity of a social media platform). The labour issues of content moderation [54,30] are displaced back onto users, and likely place further burden on already marginalised and abused people.

Contextualisation here appears as Twitter side-stepping moderation and accountability, shifting the focus onto misinformation. This is a key concern for governments, so it is something Twitter needs to be seen to be taking action on. But that means that Birdwatch is also a tactical play against the imposition of further regulation. By giving tools to communities, Birdwatch has the potential to improve context, data literacy and representation. But the hidden and conflicting motives detract from this potential to tackle specific issues with online harms.

Birdwatch could be a thoughtful way of addressing some of the complex issues around online content and harms. Emphasising context is a key part of this. But it also introduces further sociotechnical vulnerabilities that we must pay attention to. This includes: weaponisation of the platform through potential gaming and gamification of rankings; the perpetuation of dominant harmful or hateful narratives; the continuation of control by opaque algorithms even as data is made more transparent; US-centrism in establishing online norms for global communities; exploitative labour through crowd-sourcing; and the veneer of community to avoid responsibility.

Birdwatch, and the watchers of Birdwatch, will need to keep the social effects of categorisation, individualisation, exploitation, inequity, constraining discourse and responsibility for moderation in mind as the system develops.

6 Conclusion

There is a lot of potential in Birdwatch. But the issue is complex and carries significant risk of generating further vulnerabilities for public discourse. In this article, we have examined the Birdwatch pilot scheme and plans for global rollout. We have identified sociotechnical vulnerabilities under two categories: vulnerabilities within the Birdwatch system; and vulnerabilities for public discourse created by the Birdwatch system. We have outlined a taxonomy of these interrelated vulnerabilities, and discussed how mitigations for the former can lead to an increase in the latter. We have then situated these concerns in wider social issues around online communication, including issues of power and agency.

The importance of watching the design of new sociotechnical systems - and the power structures and priorities behind them - cannot be overstated. We must continue to watch the Birdwatchers, as Twitter has made possible through transparent access to data on the pilot. But we must also continue to watch Birdwatch itself. This is a more difficult task as the decision-making and algorithmic systems remain much more opaque. But it is essential that Twitter is held accountable for deciding what and how information is shared and seen on its platform. We must continue watching Birdwatch closely.

References

1. @Birdwatch. Private correspondence (Twitter DM), 29 March 2021 & 15 April 2021.
2. Allen, J., Arechar, A.A., Pennycook G., Rand, D.G.: Scaling up fact-checking using the wisdom of crowds. Preprint (2020), accessed 22 April 2021.
3. Arun, C.: Rebalancing Regulation of Speech: Hyper-Local Content on Global Web-Based Platforms. Medium - Berkman Klein Center (2018), accessed 23 April 2021.
4. Banchik, A.V.: Disappearing acts: Content moderation and emergent practices to preserve at-risk human rights-related content. New Media Soc. (2020), pp. 1-18.
5. Bell, E.J., Owen, T., Brown, P.D., Hauka, C., Rashidian, N.: The Platform Press: How Silicon Valley Reengineered Journalism. Tow Center for Digital Journalism (2017), pp. 1-105.
6. Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S.: On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. FAccT '21, pp. 610-623.
7. Benjamin, G.: From protecting to performing privacy. Journal of Sociotechnical Critique 1(1) (2020), pp. 1-30.
8. Benjamin, R. Race After Technology: Abolitionist Tools for the New Jim Code. Polity (2019).
9. Binns, R., Veale, M., van Kleek, M., Shadbolt, N.: Like trainer, like bot? Inheritance of bias in algorithmic content moderation. International conference on social informatics (2017). Springer, pp. 405-415.
10. Browne, S. Dark Matters: On the Surveillance of Blackness. Duke UP (2015).
11. Buolamwini, J., Gebru, T.: Gender shades: Intersectional accuracy disparities in commercial gender classification. FAT* '18, pp. 77-91.
12. Citron, D.K., Solove, D.J.: Privacy Harms. GWU Legal Studies 11 (2021), pp. 1-56.
13. Cobbe, J., Singh, J.: Regulating Recommending: Motivations, Considerations, and Principles. EJLT 10(3) (2019), pp. 1-49.
14. Coleman, K.: Introducing Birdwatch, a community-based approach to misinformation. Twitter Blog (2021), accessed 22 April 2021.
15. Common, M.F.: Fear the reaper: How content moderation rules are enforced on social media. Int. Rev. Law, Comput. Technol. 34(2) (2020), pp. 126-152.
16. D'Ignazio, C., Klein, L.: Data Feminism. MIT Press (2020).
17. Davidson, T., Bhattacharya, D., Weber, I.: Racial bias in hate speech and abusive language detection datasets. ACL '19 ALW3, pp. 1-11.
18. Douek, E.: The Rise of Content Cartels. Knight First Amendment Institute at Columbia (2020), pp. 1-51.
19. Douek, E.: Governing Online Speech: From 'Posts-As-Trumps' to Proportionality and Probability. Colum. L. Rev. 121(1) (2021), pp. 1-70.
20. Fan, J., Zhang A.X.: Digital Juries: A Civics-Oriented Approach to Platform Governance. CHI '20, pp. 1-14.
21. Fisher, M.: Inside Facebook's Secret Rulebook for Global Political Speech. New York Times (2018), accessed 23 April 2021.
22. Gangadharan, S.: Context, Research, Refusal: Perspectives on abstract problem-solving. Our Data Bodies (2020), accessed 21 April 2021.
23. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J.W., Wallach, H., Daumé, H., Crawford, K.: Datasheets for Datasets. FAT* '18, pp. 1-17.
24. Geiger, R.S.: Bot-based collective blocklists in Twitter: the counterpublic moderation of harassment in a networked public space. ICS 19(6) (2016), pp. 787-803.
25. Gerrard, Y.: Beyond the hashtag: Circumventing content moderation on social media. New Media Soc. 20(12) (2018), pp. 4492-4511.

26. Gillespie, T.: Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media. Yale University Press (2018).
27. Gorwa, R.: The platform governance triangle: Conceptualising the informal regulation of online content. IPR 8(2) (2019), pp. 1-22.
28. Gorwa, R., Binns, R., Katzenbach, C.: Algorithmic content moderation: Technical and political challenges in the automation of platform governance. BD&S 7(1) (2020), pp. 1-15.
29. Gorwa, R., Guilbeault, D.: Unpacking the social media bot: A typology to guide research and policy. P&I 12(2) (2020), pp. 225-248.
30. Gray, M., Suri, S.: Ghost work: How to stop Silicon Valley from building a new global underclass. Houghton Mifflin Harcourt (2019).
31. Grimmelmann, J.: The virtues of moderation. YJoLT 17(42) (2015), pp. 42–109.
32. Hamidi, F. Scheuerman, M.K., Branham, S.M.: Gender recognition or gender reductionism? The social implications of embedded gender recognition systems. CHI '18, pp. 1-13.
33. Hanna, A. & Park, T.M.: Against Scale: Provocations and Resistances to Scale Thinking. CSCW '20, pp. 1-4.
34. Hasinoff, A.A., Gibson A.D., Salehi, N.: The promise of restorative justice in addressing online harm. Brookings Institute (2020), accessed 22 April 2021.
35. Information Commissioner's Office: Letter from the Information Commissioner ICO/O/ED/L/RTL/0181. ICO (2020), accessed 23 April 2021.
36. Jhaver, S., Ghoshal, S., Bruckman, A., Gilbert, E.: Online harassment and content moderation: The case of blocklists. TOCHI 25(2) (2018), pp. 1-33.
37. Jiang, J.A., Scheuerman, M.K., Fiesler, C., Brubaker, J.R.: Understanding international perceptions of the severity of harmful content online. PLoS one 16(8) (2021), e0256762.
38. Keyes, O.: The misgendering machines: Trans/HCI implications of automatic gender recognition. CSCW '18, pp. 1-22.
39. Krasodomski-Jones, A., Judson, E., Smith, J., Miller C., Jones, E.: Warring Songs: Information Operations in the Digital Age. Demos (2019), accessed 26 April 2021.
40. Lampe C., Resnick, P.: Slash (dot) and burn: Distributed moderation in a large online conversation space. CHI '04, pp. 543–550.
41. Land, M.: Regulating Private Harms Online: Content Regulation under Human Rights Law. In Jorgensen, R.F. (ed.): Human rights in the age of platforms, pp. 285-315. MIT Press (2019).
42. Maddox, J.: Will the Gamification of Fact-Checking Work? Twitter Seems to Think So. Medium - Start It Up (2021), accessed 22 April 2021.
43. Maddox J., Malson, J.: Guidelines Without Lines, Communities Without Borders: The Marketplace of Ideas and Digital Manifest Destiny in Social Media Platform Policies. SM+S April-June (2020), pp. 1-10.
44. Marwick, A.E., boyd, d.: Networked privacy: How teenagers negotiate context in social media. NMS 16(7) (2014), pp. 1051-1067.
45. Massanari, A.: Playful Participatory Culture: Learning from Reddit. AoIR '13, pp. 1-7.
46. Matias, J.N.: The civic labor of volunteer moderators online. SM+S 5(2) (2019), pp. 1-12.
47. Micallef, N., He, B., Kumar, S., Ahamad M., Memon, N.: The Role of the Crowd in Countering Misinformation: A Case Study of COVID-19 Infodemic. IEEE BigData2020 (2020), pp. 1-10.

48. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I.D., Gebru, T.: Model cards for model reporting. FAT* '19, pp. 220-229.
49. West, S.M.: Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. NMS 20(11) (2018), pp. 4366-4383.
50. Nissenbaum, H. A contextual approach to privacy online. Daedalus 140(4) (2011), pp. 32-48.
51. Noble, S. Algorithms of Oppression. NYU Press (2018).
52. Pennycock, G., Rand, D.G.: Fighting misinformation on social media using crowdsourced judgments of news source quality. PNAS 116(7) (2019), pp. 2521-2526.
53. Reddit: Rainbow Six Siege players who use slurs are now getting instantly banned. Reddit (2018), accessed 23 April 2021.
54. Roberts, S.: Behind the Screen. Yale University Press (2019).
55. Sap, M., Card, D., Gabriel, S., Choi, Y., Smith, N.A.: The risk of racial bias in hate speech detection. ACL 57 (2019), pp. 1668-1678.
56. Suzor, N.: Understanding content moderation systems: New methods to understand internet governance at scale, over time, and across platforms. In: Ryan Whalen (editor) In Computational Legal Studies. Edward Elgar (2020). pp. 166-189.
57. Tabassi, E., Burns, K.J., Hadjimichael, M., Molina-Markham, A.D., Sexton, J.T.: A taxonomy and terminology of adversarial machine learning (draft). NIST IR 8269 (2019), accessed 21 April 2021.
58. Tufekci, Z.: Big questions for social media big data: Representativeness, validity and other methodological pitfalls. ICWSM 8(1) (2014), pp. 505-514.
59. Turkewitz, N.: The Week in Tweets: The "Scale Doesn't Scale" Edition feat. Emily Bell. Medium (2020), accessed 22 April 2021.
60. Katlyn Turner, Danielle Wood and Catherine D'Ignazio, 2021. "The Abuse and Misogynoir Playbook," In: Abishek Gupta et al. (editors), The State of AI Ethics Report January 2021. Montreal AI Ethics Institute. pp. 15-34.
61. Twitter: Birdwatch Guide (2021), accessed 7 September 2021.
62. UN Human Rights Council: Report of the independent international fact-finding mission on Myanmar. OHCHR A/HRC/39/64 (2018), accessed 23 April 2021.
63. Williams, J., Chowdhury, R.: Introducing our Responsible Machine Learning Initiative, Twitter Blog (2021), accessed 23 April 2021.
64. Wong, J.C.: How Facebook let fake engagement distort global politics: a whistleblower's account. The Guardian (2021), accessed 23 April 2021.
65. York, J.C.: Syria's Twitter spambots. The Guardian (2011), accessed 23 April 2021.
66. York, J.C., Zuckerman, E.: Moderating the Public Sphere. In: Jorgensen, R.J. (ed.): Human rights in the age of platforms, pp. 137-162. MIT Press (2019).
67. Zannettou, S.: "I Won the Election!": An Empirical Analysis of Soft Moderation Interventions on Twitter," ICWSM '21, pp. 1-13.

A Description of the Taxonomy

Key

Scale [A] Individual, [B] Type of individual, [C] Community, [D] Public discourse, [E] Systemic/principles.
Timeframe [1] Immediate, [2] Short, [3] Mid, [4] Long, [5] Persistent.

Targets

Pilot: vulnerabilities during the US and subsequent pilots
Implementation: the transition to global context;
Tweets: user interactions and vectors for attacks;
Notes: user interactions and vectors for attacks;
Note ranking: user interactions and vectors for attacks;
Twitter management: internal decisions create/mitigate vulnerabilities;
Regulation: external (legislative) decisions prevent/permit vulnerabilities.

Attacks

A 1 **Posts**: tweets and notes, including many directly abusive tactics;
A 2 **Followers**: tools in extended abuse;
A 2 **Bots**: automating abuse to scale up coordinated attacks;
A 1 **Text-as-image**: an attack on whether the Birdwatch system is triggered;
A 2 **Iterating tweets**: as above;
A 1 **Varying content or categories**: as above;
A 1 **Abusive or harmful content**: as above;
C 3 **Data**: includes data poisoning of ranking system;
B 2 **Algorithm**: gives differential visibility or credibility to tweets/notes;
C 3 **Third party**: external attacks exploit vulnerabilities;
E 4 **Design**: internal flaws create vulnerabilities;
C 3 **Injection**: includes user interactions (likely) and breached security (less so);
C 3 **Manipulation**: includes user interactions and breached security;
E 4 **Data structures**: design flaws enable manipulation of data/the algorithm;
D 4 **Faux transparency**: data availability risks obscuring underlying structures;
D 4 **External validation**: public scrutiny and PR as tool for credibility;
D 5 **Lobbying**: pressure on regulators to prevent external constraints;
D 5 **Loopholes**: flaws in regulation (e.g. loose definitions) enable harmful design;
D 5 **Self-regulation**: Birdwatch is part of continued efforts to avoid regulation.

Harms

A 2 **Coordinated attacks**: combining attacks/accounts increases scale/impact;
B 3 **Weaponisation**: systematic targeting of certain groups/communities;
A 1 **Abuse**: effects (emotional, physical) against specific individuals(/groups);
B 3 **Verification**: shift for Twitter; harms marginalised groups with need for ID;
E 4 **Policies/enforcement**: precedent of unequal application/lack of context;
C 5 **Access/exclusion**: design, policies, implementation; method & type of harm;
D 3 **Game rankings**: vulnerabilities in practice and in credibility;
D 4 **Ranking system**: vulnerabilities to public discourse in Birdwatch design;
D 4 **Avoiding moderation**: not moderation; community not platform;
D 5 **Avoiding regulation**: visible action to placate regulators;
C 3 **Exploitative labour**: reliance on users; lack of protection; uneven burden;
E 5 **Avoid scrutiny**: systemic avoidance or deflection of external audit/criticism;
E 5 **Systemic/narrative**: structural impact on society; influence over debates;
E 4 **Context shift**: marginalisation of geospatial/cultural/etc. communities.