DEMOS

A PICTURE OF HEALTH

MEASURING THE COMPARATIVE HEALTH OF ONLINE SPACES

JOSH SMITH
ALEX KRASODOMSKI-JONES
MARIA OLANIPEKUN
ELLEN JUDSON

MARCH 2021

This project was supported by GCHQ

Published by Demos January 2021

15 Whitehall, London, SW1A 2DD
T: 020 3878 3955
hello@demos.co.uk
www.demos.co.uk

CONTENTS

INTRODUCTION	PAGE 4
EXECUTIVE SUMMARY	PAGE 6
METHODOLOGY	PAGE 7
ANALYSIS	PAGE 13
CASE STUDIES	PAGE 26
CONCLUSIONS	PAGE 33

INTRODUCTION

From the early days of the web, websites have been judged by the content and communities that call them home. Thirty years ago, Usenet groups emerged, succinctly labelling their contents: soc. politics, talk.bizarre or alt.tasteless. In the decades that followed, spaces both mainstream and alternative took root, each built around a distinct set of user experiences, expectations, cultures, content types and so on. In much the same way as one might feel safe in one part of town and unsafe in another, delighted by one street and disgusted by another, at home in one place or an alien in the next, we've always known that some bits of the web are good, and others are bad. It's the bad ones that have frequently become infamous: splashed across front pages, or vainly blocked by the school firewall.

Things have changed. The rapid monopolisation of digital real estate by major platforms like Google, Facebook and YouTube has blurred these distinctions. Where in the past, avoiding the bad bit of town was easy if you knew where it was, the nature of these giant, user-driven networks means the unexpected might be lurking around the corner. On the other hand, platform-funded moderators, algorithmic prioritisation and systematised user-reporting have introduced some measure of policing for the content shared by and communities using their technology.

Although we might feel we know good from bad when we see it, much less attention has been given to actually trying to measure the health of online spaces. The metrics we have are often clunky, opaque or misapplied. We have scant understanding of why certain spaces attract, promote or condone certain behaviours, and the relative importance of factors like culture, moderation practices, architecture, history or users.

We do this much better in the offline world. In the UK, the Thriving Place Index (TPI) is a multi-dimensional measure of community health covering all upper and second tier Local Authorities in England and Wales.[1] It brings together indicators from a range of bodies, including the Office for National Statistics (ONS), Public Health England and the Index of Multiple Deprivation. It covers both direct measures and proxies for factors known to have an influence on wellbeing. Local conditions are split further into an array of subcategories. In addition to various objective indicators (such as household income), the TPI currently includes various subjective measures, including self-reported disability, self-reported state of health, perception of neighbourhood trust levels, and perception of safety after dark.

The ONS also carries a category for measures reflecting the health of our offline spaces: "an individual's dwelling, their local environment and the type of community in which they live".[2] It collates crime, perception of safety after dark, access to nature, feeling a sense of belonging to one's neighbourhood, travel time to key services and satisfaction with accommodation. Other measures attempt to understand how subjective feeling of the wellbeing of a place connects to particular issues.

And we know how to use features of offline places to infer how certain harms may be more likely to occur there. HOPE not hate's recent report, *Understanding Community Resilience in our Towns*, attempts to understand which towns are receptive to divisive politics.[3] It clusters 862 towns using over 100 variables for each, including more widely-relevant measures such as indices of multiple deprivation and ONS economic data, alongside measures such as far-right activity and segmentation data on attitudes regarding immigration and religion.

This is all to say that we have a sophisticated framework for understanding the health of offline spaces. It is now time to systematise the measurement of the health of online spaces. There are already existing tools, technologies and methodologies to help do this, developed by industry, academics and think tanks. Governments

1. The Thiving Places Index. Available at https://www.thrivingplacesindex.org/page/about/measurement (accessed December 2020)
2. Measures of National Well-being Dashboard, ONS. Available at https://www.ons.gov.uk/peoplepopulationandcommunity/wellbeing/articles/measuresofnationalwellbeingdashboard/2018-04-25 (accessed December 2020)
3. Understanding Community Resilience in our Towns, August 2020. Available at http://www.hopenothate.org.uk/wp-content/uploads/2020/09/Understanding-community-resilience-in-our-towns.pdf (accessed December 2020)

ought to build on these, and use this evidence to better understand the online worlds their citizens are increasingly living in. And this evidence should be used to inform regulation, to ensure that the forces and structures responsible for the gap between healthy and toxic are taken into account. This paper presents a short study in highlighting some of these forces.

EXECUTIVE SUMMARY

- Analysts algorithmically classified comments for toxicity from a dataset drawn from 80 subreddits on the platform Reddit and explored the ways in which the users, norms and structures of these spaces contributed to their relative toxicity.
- Analysts found a meaningful difference between subreddits when judged by overall toxicity. Natural language processing (NLP) algorithms are cost-effective and efficient routes for platforms to judge the contents of their spaces, but given the limitations of algorithmic classification in judging context their use should be carefully considered.
- However, average toxicity as an absolute measure only partially characterises a space. It fails to differentiate instances of highly toxic behaviour from a generally more toxic environment, and fails to take into account the extent to which a toxic comment might diverge from the norms of a space.
- Larger spaces contain a higher proportion of highly toxic comments.
- Broader rule sets are associated with positive adjustments in toxicity.
- Some rules appear to have a stronger impact on the average toxicity of a space than others.
- The number of moderators present in a space had no effect on our toxicity measures, though all spaces had at least some moderation.
- The number of comments made by a user used as an approximation for their relative activity in a space had no effect on the average toxicity of the user's comments.
- Importantly, a user's behaviour does change in toxicity across different online spaces, but the effect is only extreme in a minority of cases.
- Users in online spaces do not behave consistently across multiple spaces, but adjust their toxicity or civility to the norms, rules or cultures of the spaces they use.
- These findings have major implications for how we think about toxicity online, and the wider online harms debate. Moving the focus on to how online spaces are designed, managed, and on to the cultures that develop within them, may be more valuable than a narrow focus on aberrant behavior by individuals.

METHODOLOGY

REDDIT DATA

As a platform which allows people to create and maintain communities, Reddit is fairly liberal in allowing moderators and participants to decide how they would like that community to be run. Each subreddit must abide by a universal content policy, which, for example, prohibits threats of violence.[24] Beyond these minimal rules, however, subreddits are free to decide where to draw their own limits for acceptable content; to decide who should moderate a space, and the extent of their powers. In Jonathan Zittrain's terms, it is a "socially generative" system, in that it permits moderators to experiment with the rules governing interaction in their communities. This generativity is part of what makes Reddit such an interesting and diverse platform, as well as explaining why it has often courted controversy.[5]

This system creates spaces which can be remarkably consistent in their tone. A 2019 paper by Rajadesingan et. al found that many political subreddits were able to maintain social norms despite a high turnover of users.[6] This work showed that the strongest factor in getting newcomers to correspond to a community's norms are those which occur 'pre-entry' - for example, users reading posted rules, or observing the behaviour of others, before making a first post or comment. In the analysis below, we explore the connections between the choices made by the creators of a subreddit, and the extent to which users are likely to conform to the norms of a space.

To do this, researchers built a dataset of posts and comments from the platform Reddit, gathered through the Reddit API.[7] Analysts selected 80 subreddits: 50 popular subreddits in the UK selected to reflect a range of topics, and 30 selected subreddits supporting case studies of spaces dedicated to similar topics with seemingly contrasting levels of health. This dataset contained 24,000 posts and 350,000 comments, all made in 2020. A list of subreddits selected for the study are contained in Appendix 1 below. They can be broadly categorized as:

General

Popular subreddits dedicated to a range of topics. The list included subreddits devoted to news sharing, hobbies, games, left-wing and right-wing political positions, science, history and digital culture. Subreddits included r/todayilearned, r/conspiracy, r/sports and r/worldnews, as well as deeply oppositional political subreddits like r/enlightenedcentrism, r/prolife and r/toiletpaperusa.

Case Study: Covid-19

A list of subreddits emerging in the wake of the coronavirus pandemic discussing the news, supporting or protesting the governmental response, and acting as support groups for those suffering. Subreddits included r/coronavirus, r/lockdownskepticism and r/covid19_support.

Case Study: Portland Protests and Police Violence in the US

A list of subreddits discussing the violence and protests against police brutality in Portland in the summer of 2020, aimed at capturing both sides of the debate and including geographically-specific voices. Subreddits included r/bad_cop_no_donut, r/good_cop_free_donut, r/portland, r/protectandserve and r/blacklivesmatter.

Case Study: Magic: The Gathering

A list of subreddits devoted to the trading card game Magic: the Gathering. Subreddits included r/magicTCG, r/mtgfinance and r/magicarena.

4. Reddit Content Policy. Available at https://www.redditinc.com/policies/content-policy. (Last accessed December 2020)
5. For an engaging exploration of some of these controversies, as well as an illustration of the platform's excentricities, see Carl Miller's piece 'Reddit Run'. Available at https://www.the-tls.co.uk/articles/reddit-run-carl-miller/ (Accessed November 2020)
6. Rajadesingan et al., Quick, Community-Specific Learning: How Distinctive Toxicity Norms Are Maintained in Political Subreddits (2020)
7. Reddit API Documentation, https://www.reddit.com/dev/api/ (Accessed November 2020)

TOXICITY CLASSIFICATION

Determining the health of an online space is difficult: different spaces and communities online have different social norms and values, and so identifying a common indicator of health poses a methodological challenge. The aim of this research was to be able to compare moderation practices in spaces which court controversy as much as those which tightly control it. Once comments were collected, they were classified by toxicity using a natural language processing (NLP) algorithm.

To measure toxicity, we use the Jigsaw Perspectives API, developed in collaboration between Google and the Wikimedia Foundation.[8] While this API was trained on Wikipedia, it has been effectively applied by previous researchers to Reddit data, and in tests (described below) was found to perform well on our dataset.[9] The Perspectives API predicts the perceived impact a comment may have on a conversation by evaluating that comment across a range of emotional concepts, called attributes.[10] The toxicity attribute is defined as "a rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion". Each comment was sent to the API and scored for toxicity on a scale of 0 (low toxicity) to 1 (high toxicity). The score reflects the overall summary score for the whole body of text and is indicative of the probability that a reader would perceive the comment as being toxic. The API is only able to analyse text - therefore any results solely containing emojis or other non-word characters were omitted.

A toxicity score alone can be a misleading metric as it tends to exclude context from a comment. The reclamation of slurs by minority groups and LGBTQ+ groups, for instance, has confused algorithmic content moderation systems which view this speech as broadly indistinguishable from slurs themselves.[11] In short, what may be seen as toxic in one space may be perfectly acceptable in another. The following descriptions of toxicity bands adds useful context, however, and as part of this research we do draw a line: comments scoring higher than 0.8 were deemed highly toxic for the contexts we examined. A sample of comments were checked to ensure the classification was working accurately. Toxicity scores were rounded, and bands are described below.

Friendly debate: toxicity scores of 0.0 - 0.19 (55% of comments)

Comments in this category fell into two broad categories: friendly encouragement, debate or agreement, or content that contained little to no emotive language at all, such as providing a link to a third-party website or stating a fact without the provision of any opinion of another Reddit user's contribution.

> *Interesting story, and I definitely agree with your last sentence there.*
>
> *Cash is anonymous too. So are checks*
>
> *Question: How does everyone's computer run Warcraft (if you play it)? Not as many things on board, but some of the animations are pretty amazing.*

Respectful disagreement: toxicity scores of 0.2-0.39 (21% of comments)

Comments in this category expressed an opinion, usually in language that was respectful of other users.

> *It's very annoying when a comment like this where the person doesn't know what they're talking about gets so popular*
>
> *I'd argue that's the point: social unrest and constant division. Perfect for an election year!*
>
> *I would also note Korea is pretty xenophobic despite them being relatively well-off.*

Polite but angry: toxicity scores of 0.4-0.59 (9% of comments)

Comments in this band remained broadly civil, but were more likely to be mocking of other opinions. Comments expressing anger, disbelief and disappointment tended to group in this band of toxicity scores. Language remained polite for the most part, but included a range of more radical beliefs identified by the Perspectives API as being more likely to provoke a negative reaction.

> *Did you look at the URL? I'm outraged and pissed. Also I didn't hear about ICE with the 1488. I'm so angry*
>
> *I'm 23 and have been over this shit since the beginning of April. I'm embarrassed that so*

8. See Wulczyn, Thain and Dixon, Ex Machina: Personal Attacks Seen at Scale (2020).
9. See Rajadesingan et al., Quick, Community-Specific Learning: How Distinctive Toxicity Norms Are Maintained in Political Subreddits (2020) and Mittos et. al "And We Will Fight For Our Race!" A Measurement Study of Genetic Testing Conversations on Reddit and 4chan" (2019)
10. For more documentation, and to use this resource yourself (which was free at time of writing), see Perspectives API: https://support.perspectiveapi.com/s/about-the-api
11. Bours, B. Facebook's Hate Speech Policies Censor Marginalized Users. Wired, 2019. Available at https://www.wired.com/story/facebooks-hate-speech-policies-censor-marginalized-users/ [accessed 02.02.21]

many people my age are content with just putting their lives on hold forever. I want to finish up college (the normal way), meet new people, go to new places, etc. It's been almost five months, and I swear sometimes it's gotten worse than the March hysteria. The hypocrisy of people truly amazes me as well. Like I know it's usually fake when someone claims to care about others (a truly good person doesn't need to tell others they're good), but it's still jarring when the ones you're close to turn into completely different people.

Defunding the police isnt stupid. Its reducing the many jobs we've forced the police to take over, funding the appropriate departments to hire professionals that can actually deal with those problems, and having police focus only on violent crime. Bringing up the house burning is just as much propaganda as anything OP said. Check your own glass house before throwing rocks.

Courting controversy: toxicity scores of 0.6-0.79 (8% of comments)

Comments here tended to touch on controversial subjects as well as using obscenities and offensive speech, usually to emphasise a point. Nevertheless, comments tended to avoid attacking other Reddit users through insults, though most often by questioning their suitability or ability to participate in the discussion or debate.

Again, boomers (largest generation and highest % voters) control US politics. Politicians enacted the lockdowns. Lockdowns disproportionately fuck over young people.

*And the bootlicking shitheels from r/*** cheering it on makes this worse. Disgusting filth. The cops and anyone who thinks its ok. That sub seems pretty bloodthirsty all around. They seem to salivate over police brutality and violent vigilante justice. What a vile sub.*

And the majority of people agreed with me. Regardless, its cathartic to lay out the facts to sub 80 IQ liberals like you that live in a dream world. I can link you the video of facebook openly discriminating against White males and conservatives in hiring and you will just respond with typical liberal idiocy. Keep laughing soyboy, everyday more people are waking up to your lies that you and CNN peddle.

Toxic violence: toxicity scores of 0.8-1.0 (7% of comments)

These comments were the most toxic, including speech easily classed as racist and anti-semitic. Comments assigned this toxicity were frequently targeting other Reddit users with accusations or demands, did not contribute to any constructive debate, and tended to contain the highest number of obscenities.

how fucking retarded you sound. you're too retarded to realize.

"Feminism is about women wanting to be independent beings that can do what they want, supposedly" Holy shit, what a fucking idiot

all bankers deserve an excruiccccciating death, they're all fucking bloodsuckers

It is worth noting that text-based analytics are not perfect. Typos or unfamiliar language can trip the classifier up, and a 100% accuracy is impossible. Gauging the relative toxicity of comments on a spectrum would be challenging even for humans to do. In our experience, we would expect a classifier attempting to gauge toxicity to agree with a team of human analysts approximately 70% of the time.

Analysts carried out a check of 200 classified comments to help inform our view of the classifier's accuracy. In this check, the analyst and algorithm agreed 78% of the time, suggesting the classifier is performing well. It is noteworthy that this algorithm is available to all social researchers, and we would encourage others to test its accuracy further.

The Perspectives API has received some criticism for assigning greater toxicity to comments which are affirming or empowering but which e.g. use swear words in doing so, than to comments which denigrate groups but may do so in a different tone.[12] To some extent, this is likely to be a product of 'toxic' as defining comments which seem likely to stop discussion rather than comments which are morally offensive, combined with the inevitable lack of 100% accuracy from an algorithmic assessment

12. Drag Queen vs. David Duke: Whose Tweets Are More 'Toxic'?, Wired (2019). Available at https://www.wired.com/story/drag-queens-vs-far-right-toxic-tweets/ (Last accessed December 2020)

of content. However, it does show that platforms relying on toxicity alone to measure the health or quality of discussion in a space would be insufficient and would risk engaging in 'tone policing' rather than genuine empowerment, promotion of good discussions, and reduction of violence. Thus though we argue that assessments of a platform's health should be supported using methods and tools like using NLP classifiers to assess toxicity, it should always be as part of a suite of measures and with strong oversight of how those algorithms are being deployed and their impact.

GENERATIVITY IN EVIDENCE: DIFFERENT APPROACHES TO MANAGING SUBREDDITS

In this report, we aim to measure whether the characteristics of a space have an effect on the 'health' of that space - a term we more tightly define below. Put simply, one question is why despite being nominally part of one platform, some subreddits display very low toxicity scores and others score more highly. The second question is whether that is a function of the users in that space, or the rules, cultures and architecture of those spaces themselves. To put it simply: are toxic users limiting their toxic activity to toxic spaces, or do they take their toxic behaviour everywhere?

Rajadesignan et al. suggest that a new user in a space is most likely to be influenced by what they term 'pre-entry learning' - the processes by which a user works out what is likely to be acceptable in a subreddit, for example, through reading the conversations already underway, observing which comments are being upvoted, or checking the community's rules - something which numerous subreddits encourage users to do on their front pages. Rajadesingan shows that pre-entry learning contributes to subreddits maintaining their toxicity levels over time. The paper does not, however, explore what it is about a subreddit which has the strongest effect on convincing newcomers to conform. Below we examine three characteristics chosen to help explore this question: the number of rules, the number of moderators, and the size of the space.

Two of these characteristics concern the architecture of a space, as designed by its creators and controllers - the number of moderators who are assigned with protecting each subreddit, and the extent to which a subreddit is governed by explicit rules, which we have broken into separate categories below. The third is the size of the community - the number of different users posting on the subreddit. Each characteristic represents an aspect of a subreddit which users might observe when deciding how to participate in a conversation.

Rules

To measure the rules present in each subreddit, analysts reviewed the visible rules pertaining to each subreddit. A subreddit is free to set its own rules that sit alongside the overall terms of service of the platform, and tend to either strengthen and re emphasise those sitewide rules or set greater restrictions or expectations on the types of behaviour permitted in the space.

Rules were coded through grounded theory and grouped. The rules identified are shown below, with examples taken from the data. Rules were divided into the following categories:

Post Quality Rules (41 Subreddits)

> *"All content must be original and unique.", "No Low-Effort Posts & Memes", "Stories must be plausible".*

Post Structure Rules (44 Subreddits)

> *"No forbidden titles, no titles or posts involving Cake Days or upvotes", "No Editorialized or Misleading Titles", "Posts must be requests for advice OR clarification".*

No Personal Information/Nothing Personal (33 Subreddits)

> *"Don't Address Individuals or Distinct Entities", "Blur out identifying info (Age, Name(s), Location, etc). This includes OP's info.", "No personal opinions/anecdotes/ subjective posts"*

Respect/Civil Behaviour (53 Subreddits)

> *"Respectability", "Please be nice to other users", "No Uncivil Comments"*

No Bigotry/Racism/Sexism/Hate/Offensive Content (50 Subreddits)

> *"No Bigotry or Offensive Content", "Don't be hateful, insensitive, or inappropriate, and no offensive slurs", "Use of language which perpetuates any form of destructive hierarchy such as sexism, ableism, transphobia, homophobia, racism, etc."*

Stay On-Topic (41 Subreddits)

> *"[No] Off-topic: Not explicitly about US politics", "Common Reposts + Clutter", "No Repost/Spam"*

No Spam/Reposts (43 Subreddits)

> *"[No] Spam / Excessive self-promotion", "No reposts", "No clickbait, blogspam, or self-promotion"*

No Politics (12 Subreddits)

> *"No political or politics related tips", "No US Internal News or Politics", "No politics or political figures"*

No Memes (25 Subreddits)

> *"No memes or memetic content of any kind", "[No] Memes, Reaction, "Top 10" Lists, Blog Posts", "No Memes, Gifs, unlabeled NSFW images"*

No Sales/Commercial Posts (16 Subreddits)

> *"Belongs in the Classifieds thread", "No advertising, fundraising, surveys or studies", "No Solicitations for donations"*

Limited Posting Rights (6 Subreddits)

> *"If you are new, you'll need to build up your karma in this fashion: comments first -> then link submissions -> then text submissions", "Submission Statements are required for link and image posts. Posts w/o SS are removed after 20 min.", "Posting rights require minimum qualifications"*

Moderators have Final Say (10 Subreddits)

> *"Mods may remove/ban for any reason", "Moderators have the final say", "User history/Moderator Discretion"*

Exclusionary to a Group (8 Subreddits)

> *"No SJW/LGBT Proselytizing", "Fascists. They will be banned", "Posts from black people only"*

These characteristics highlight genuine differences between our subreddits. As seen in the graphs below, our collection contains subreddits at each end of the spectrum for size, stated rules and the size of the moderation team. While the size of a subreddit may not be a result of the choices made by those who run it, the range seen in the

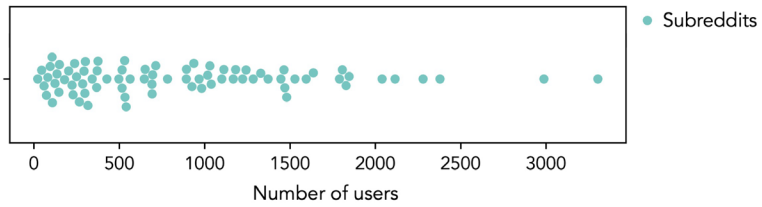


FIGURE 1.
FREQUENCY OF SUBREDDIT BY SIZE

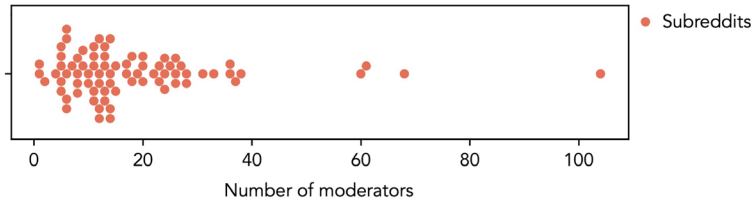


FIGURE 2.
FREQUENCY OF SUBREDDIT BY NUMBER OF MODERATORS PRESENT. NOTE THAT FOR LEGIBILITY TWO OUTLIERS ARE NOT SHOWN HERE - /R/SCIENCE (1547 MODERATORS) AND /R/ASKSCIENCE (425 MODERATORS)

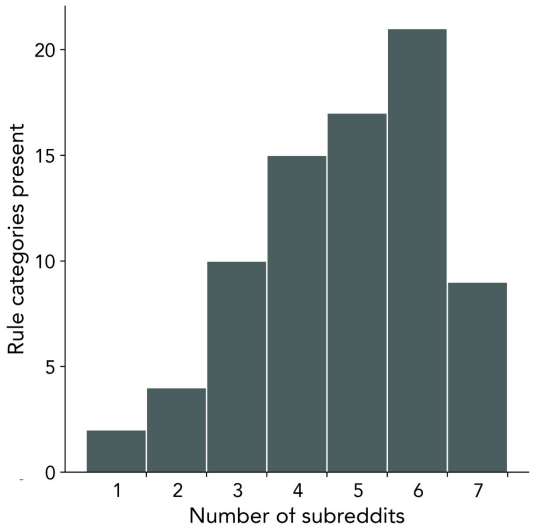


FIGURE 3.
FREQUENCY OF SUBREDDIT BY NUMBER OF RULES CATEGORY

other two measures is an example of Reddit's social generativity at work - showing that different subreddits have taken different decisions about how to govern themselves.

Taken together, the data collected reflects the broad array of different spaces found on Reddit: from large, well-moderated spaces with clear and thorough rulesets through to smaller forums with a skeleton staff and a more permissive set of behavioural standards. Comparing the behaviour within and across these differing spaces forms the core of this paper.

ANALYSIS

OVERALL TOXICITY OF THE ANALYSED SPACES

Finding One: There is a meaningful difference between subreddits when judged by overall toxicity

Analysts began with exercises to characterise the dataset as a whole, allowing us to test the applicability of the toxicity score to a real-life situation as a measure of health, and to provide an overall view of the ecosystem under investigation. Below we show the distribution of subreddits by the average toxicity score of their comments, alongside a table showing the scores for each individual subreddit.

This figure clearly illustrates the point that different subreddits operate at different levels of toxicity - there is no universal level considered acceptable. To measure whether the members of a subreddit are likely to consider a comment acceptable, it is necessary to look at the toxicity of comments in light of that community's general behaviour. Taken together, this measure acts as a baseline for each space, and a strong indicator of the kind of language and behaviour a space will permit.

The table on the following page shows the average toxicity for each of the subreddits we were able to collect data from, as well as the proportion of comments made that were classified at 0.8 toxicity or higher - the highest band described in the methodology section above.

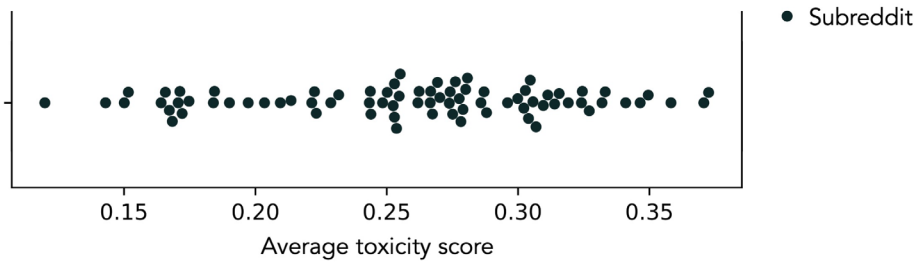


FIGURE 4.
SUBREDDITS BY AVG. TOXICITY SCORE

Subreddit	Average Toxicity	% Comments at 0.8 toxicity or higher
All subreddits	0.23	6.5%
asktrp	0.37	15.0%
trueunpopularopinion	0.37	11.0%
bad_cop_no_donut	0.36	11.8%
enlightenedcentrism	0.35	10.3%
subredditcancer	0.35	12.3%
rage	0.34	12.9%
freemagic	0.33	12.7%
ccp_virus	0.33	11.3%
blackpeopletwitter	0.33	10.9%
toiletpaperusa	0.32	11.5%
coronaviruscirclejerk	0.32	10.5%
trumpvirus	0.32	9.9%
mensrights	0.32	7.5%
breadtube	0.31	9.1%
liberal	0.31	7.4%
conspiracy	0.31	9.2%
conservative	0.31	7.0%
completeanarchy	0.31	9.5%
nottheonion	0.30	9.3%
libertarian	0.30	7.9%
nosleep	0.30	10.7%
kotakuinaction	0.30	7.8%
politics	0.30	8.0%
hong_kong	0.30	7.3%
easternsunrising	0.29	4.4%
blacklivesmatter	0.29	6.3%
aboringdystopia	0.29	9.0%
relationship_advice	0.28	7.7%
sports	0.28	10.0%
portland	0.28	8.0%
darkenlightenment	0.28	4.3%
china	0.28	5.3%
funny	0.28	10.8%
coronavirusfos	0.28	7.6%
sino	0.27	4.7%
nonewnormal	0.27	6.8%
actualconspiracies	0.27	5.8%
choosingbeggars	0.27	8.9%
protectandserve	0.27	6.6%
prolife	0.27	3.1%
military	0.27	8.4%
worldnews	0.26	5.6%
skeptic	0.26	5.2%
news	0.26	4.5%
todayilearned	0.25	7.1%
movies	0.25	6.4%
mademesmile	0.25	6.6%
good_cop_free_donut	0.25	5.3%
askreddit	0.25	7.4%
coronavirus	0.25	6.4%
hongkong	0.25	6.2%
lockdownskepticism	0.24	4.2%
collapse	0.24	4.8%
talesfromthesquadcar	0.24	5.8%
getmotivated	0.23	6.5%
magicthecirclejerking	0.23	5.9%
lifeprotips	0.22	6.2%
gaming	0.22	6.5%
china_flu	0.22	4.2%
redpillwomen	0.21	1.8%
showerthoughts	0.21	6.3%
leagueoflegends	0.20	3.5%
pcmasterrace	0.20	4.3%
tattoos	0.19	3.5%
history	0.18	1.4%
mtgfinance	0.18	3.5%
explainlikeimfive	0.17	3.5%
fitness	0.17	2.8%
science	0.17	1.2%
magicarena	0.17	2.3%
covid19_support	0.17	1.4%
spikes	0.17	1.3%
magictcg	0.17	2.3%
askportland	0.16	1.5%
travel	0.15	1.5%
covid19positive	0.15	1.2%
pewdiepiesubmissions	0.14	0.0%
askscience	0.12	0.2%

FIGURE 5.
SUBREDDITS BY AVG. TOXICITY SCORE AND PROPORTION OF HIGHLY TOXIC COMMENTS

This simple measure emerges as a useful tool to set expectations for an unknown online space. Comments in /r/travel or /r/askscience score significantly below average on toxicity scores, while Ask the Red Pill (/r/asktrp), /r/rage and /r/enlightened centrism are spaces that tolerate a higher level of toxic content. Policy and regulatory debate will continue to shift around the question of how far this kind of analysis should inform decision-making about a space: at its most crude, it might be said that toxic spaces should be under closer scrutiny than those able to prove they are more civil. On the other hand, we might say that users have an expectation that some spaces will be more tolerant of toxic behaviour and that they ought to have the freedom to participate in these should they understand the risks. For the purposes of this paper, however, it is important to note: we are able to use this analysis to sketch out the variety of cultures in our data, and the probable expectations of the users operating in them as to what is or isn't okay.

USING TOXICITY TO MEASURE THE HEALTH OF A SUBREDDIT

Finding Two: Absolute measures of toxicity only partially characterise a space. It fails to differentiate instances of highly toxic behaviour from a generally more toxic environment, and fails to take into account the extent to which a toxic comment might diverge from the norms of a space.

To measure the health of that space, then, we need to go beyond the average toxicity of its comments. The diverse social norms and subject matters discussed on the platform, and more generally across the internet, mean that a comment which would be considered a helpful addition to the debate in one space might derail the conversation in another. This is the challenge in studying a socially generative platform; any measure applied across spaces must take into account the differences between those spaces.

Below, we outline three separate measures, applied to our dataset in order to determine how the comments within each compared to the norms within that space. The questions we ask of each space are:

What proportion of the conversation is highly toxic?
Measuring proportion of comments at 0.8 toxicity or higher

How provocative are the provocateurs?
Measuring the range of the third quartile of comments by toxicity

How far will people adjust to participate?
Measuring the relative toxicity of a user across multiple spaces

The reasoning behind our use of these measures, along with what they might tell us about the health of a subreddit, is outlined in detail below.

These metrics are later tested against independent variables, such as the number of moderators a space has, or the rules they display, to see whether these have an effect on the overall health of the subreddit - at least by the above rough definition of 'health'. Our implementation of each of these metrics, and some examples showing the reasoning behind them, is explained in more detail below.

What proportion of the conversation is highly toxic?

As mentioned above, 'highly toxic' posts scoring over 0.8 were found to be likely to have cleared a threshold for speech which was likely to be problematic in all of our studied subreddits. To measure this we calculated the proportion of the comments in a space which met this high threshold for toxicity.

How provocative are the provocateurs?

This metric looks at the 25% of 'provocative' comments which were above the average toxicity for that space, to establish how far those who were willing to be more toxic than their peers were willing to go. Hypothetically, a larger range for this third quartile reflects a community without a clear sense of how toxic - or, to call back to the Perspective API's definition, how "rude, disrespectful and unreasonable" - users are expected to be in that space.

As we have seen, the average toxicity varies between subreddits. This measure looks at how comments are distributed around this average. In particular, how those people who are willing to be more toxic than the average are prepared to act in each space. Figure 6 below shows the distribution of comments made on /r/MagicTCG - one of the least toxic spaces, on average, in our dataset. This median divides the dataset in half - half of the comments collected were of a higher toxicity, and half of a lower toxicity.

The box plot below shows how people are clustered around the median. A quarter of the comments on this subreddit are contained within the shaded area on the left of the median, and another quarter on the area to its right. In this case, the box plot shows us what we can see from the histogram above it - most comments on /r/MagicTCG tend to sit in a small range, around a low toxicity. While there are some comments at the very high end of the graph, the majority of comments stick to a similar tone. This is evidence of a community which has developed norms of communication and is sticking to them.

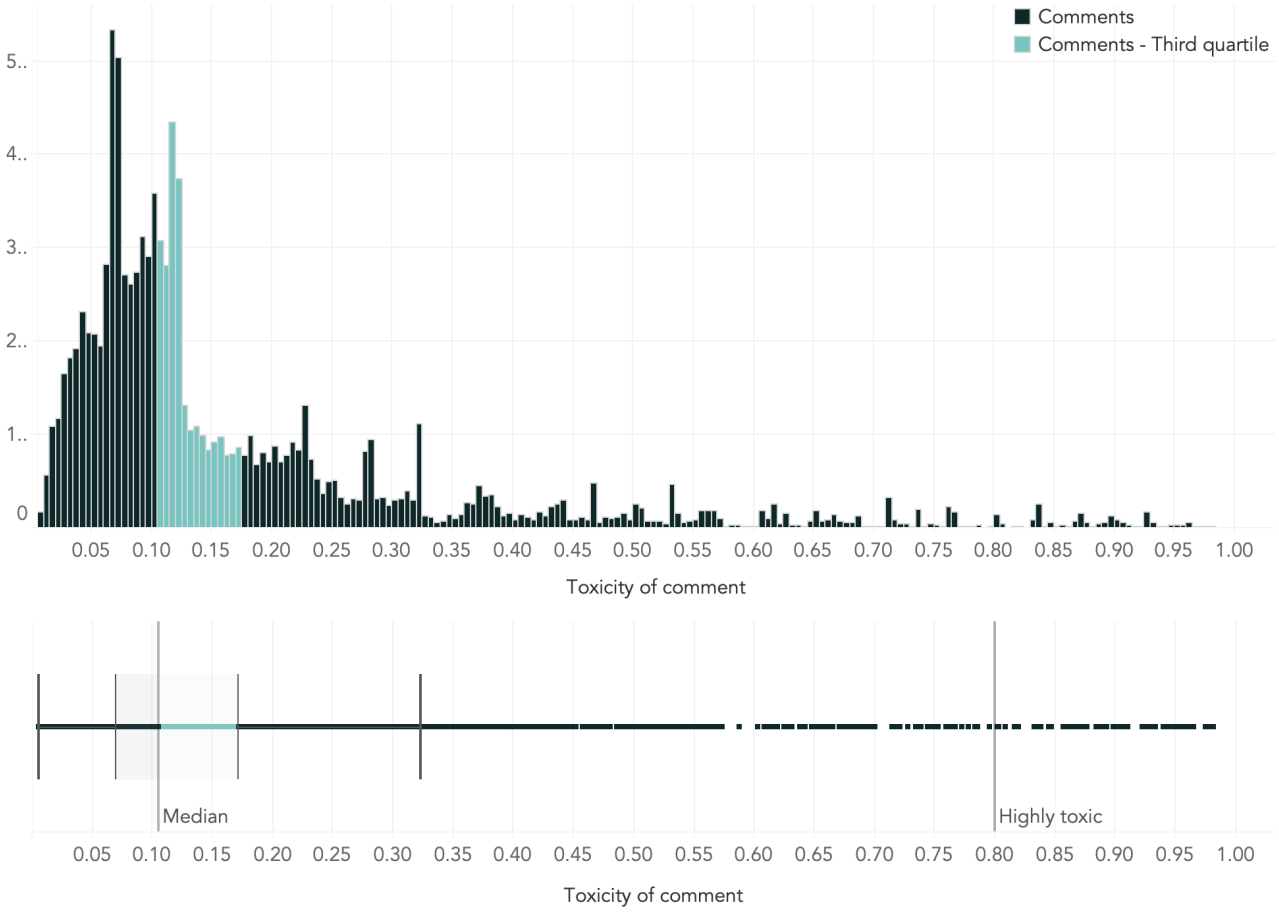


FIGURE 6.
DISTRIBUTION OF COMMENTS BY TOXICITY ON THE SUBREDDIT /R/MAGICTCG

The light blue segments represent the third quartile: that 25% of the dataset whose comments are higher than the median value of toxicity. These are the comments which are close enough to the average to be seen as within normal discussion for that space - they are not outliers, but score higher than most comments for toxicity. We'll call these 'provocative' comments. Our first proxy for the health of a community is how far this third quartile stretches along the toxicity level - i.e. how much more toxic than the average are that community's provocative comments. To illustrate this, Fig 7 below shows the same set of graphs for /r/asktrp, the most toxic subreddit in our collection. This subreddit has a higher median than /r/MagicTCG, but also a more extensive set of provocative comments; the third quartile of comments - shown in blue - stretches over a much wider range of toxicity.

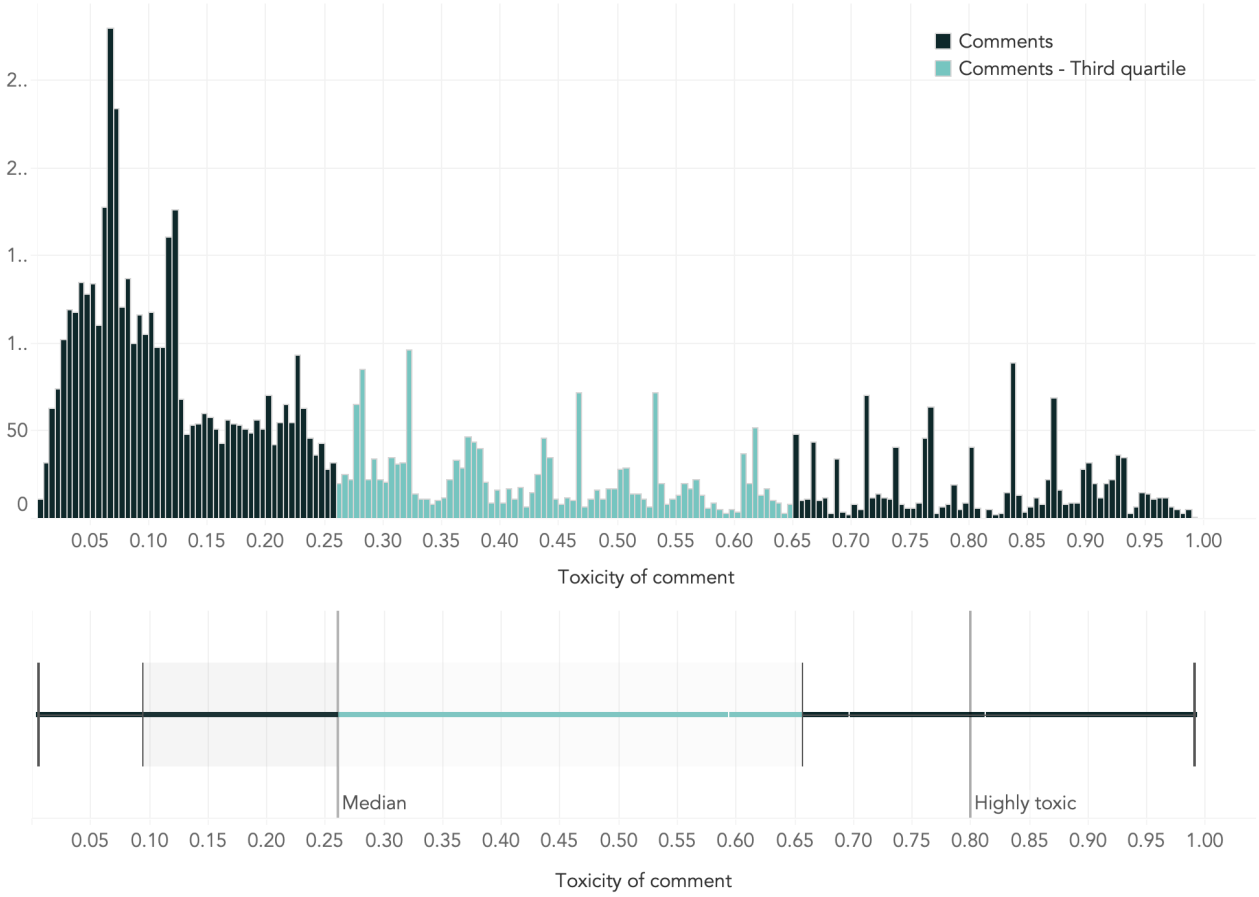


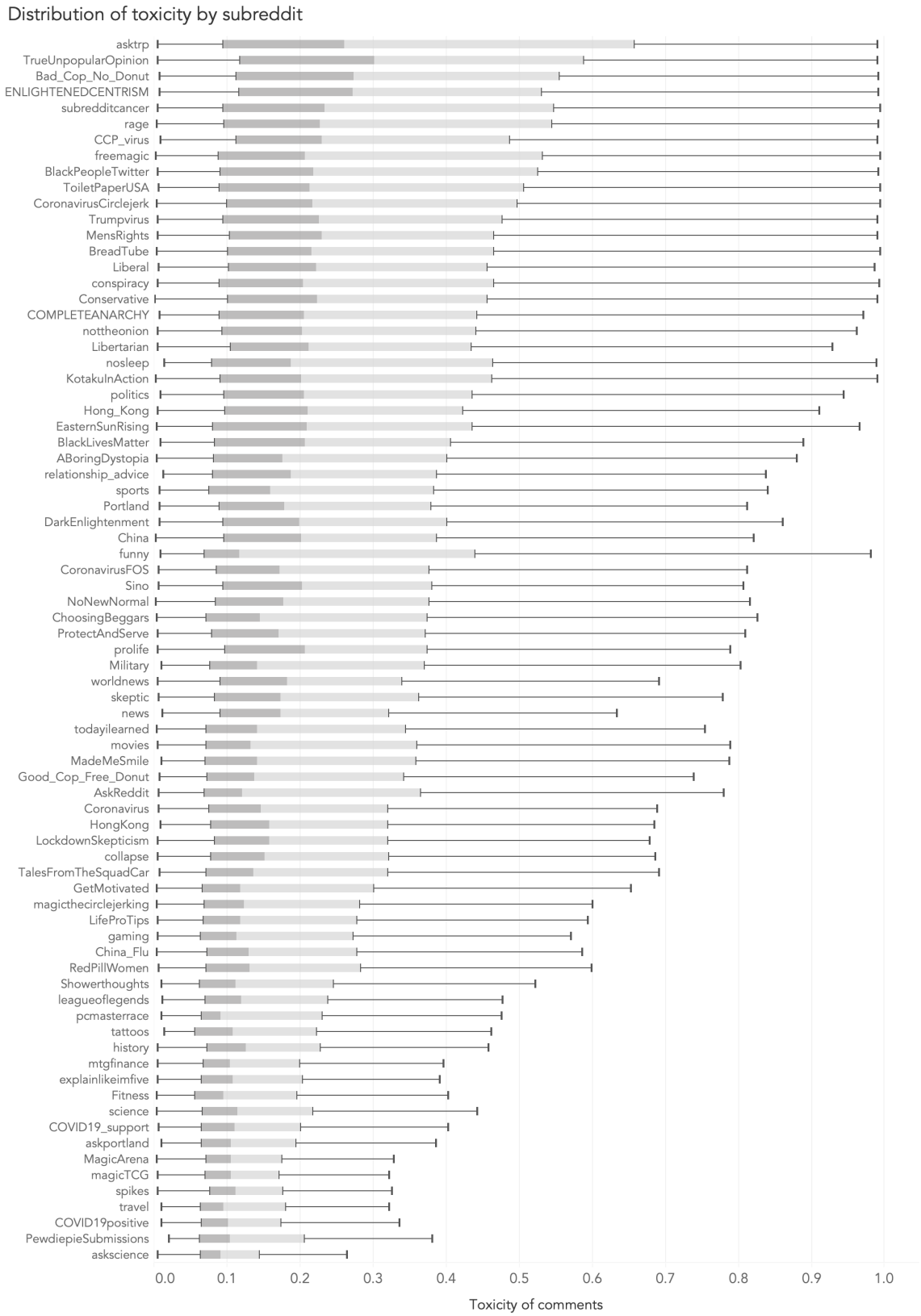
FIGURE 7.
DISTRIBUTION OF COMMENTS BY TOXICITY ON /R/ASKTRP

This measure is designed to tell us how well aligned commenters are to the norms of that space, here approximated by the average comment toxicity. Our hypothesis is that, in subreddits with strong norms, commentators will have a stronger sense of how toxic they can be in that space, and the blue quartile of 'provocative' comments will have a lower range. In spaces where these norms are less well established, this should be reflected in a wider spread of provocative comments, and a broader Q3.

The graph on the following page shows a similar boxplot for each subreddit in our collection, ranked by the range of their third quartile - or, if you like, by how provocative their provocateurs are willing to be. This illustrates that there is no straightforward relationship between this measure and the median toxicity of a space. It is possible for users to be tightly clustered around a higher average toxicity - for example, subreddits such as /r/news have a relatively high median toxicity, but appear further down this ranking due to a lower Q3 range. In contrast, /r/funny has a low median toxicity, but a broad Q3 - people in /r/funny who are willing to push the boundaries in that space will deviate further from the average than those in /r/news. This could be a result of each space's subject matter - comments discussing news are more likely to be left in the form of a discussion, and thus appear less "rude, unhelpful and disrespectful" than those intended as jokes. It is likely that this measure will be most useful in comparing subreddits with a similar subject matter. However, it is also likely to give us one indication of whether commenters are observing any social norms in discussion, and the strength of this effect.

Using this range of toxicity as a proxy for an established set of norms and cultures determining the behaviour of users in a space, with a narrow range of toxicity suggesting strong community norms and a wider range suggesting a volatile or immature set of norms, On Reddit, we see sets of strong community norms across multiple levels of toxicity. This has implications for governance, suggesting that a 'one size fits all' approach to moderating socially generative platforms is likely to fail. The tendency in popular discourse is to talk about platforms as a whole – to discuss the problems of 'Facebook'; of 'Reddit' or of 'YouTube' and demand top-down, broadbrush solutions to content moderation problems that apply equally across all spaces. This will be necessary in some cases - but what this analysis shows is that very different cultures exist within these spaces, and in some cases, a more targeted approach may be appropriate.

FIGURE 8.
DISTRIBUTION OF COMMENTS BY TOXICITY ACROSS ALL COLLECTED SUBREDDITS



How far will people adjust to participate?

We wanted to investigate whether spaces prompted people to act differently on joining them, and which spaces were effective in bringing people's toxicity down to their average level. Accordingly, this measure examines, for each subreddit, the amount by which users change their toxicity when commenting in that space, compared to their comments in the rest of Reddit. This is designed to be a positive indicator - subreddits on which users are observed adjusting their behaviour are more likely to be able to maintain a consistent tone, and will hypothetically be more constructive spaces.

One of the interesting aspects of Reddit as a collection of communities is that individuals are free to move amongst, and participate in, communities with differing social norms. We have found that users adapt their toxicity levels according to the subreddit they are commenting in. An example can be seen in the posting history of one user, displayed below. Each of the graphs here show the same data - the toxicity of each comment left by a single user in three subreddits - two political communities and a gaming forum.

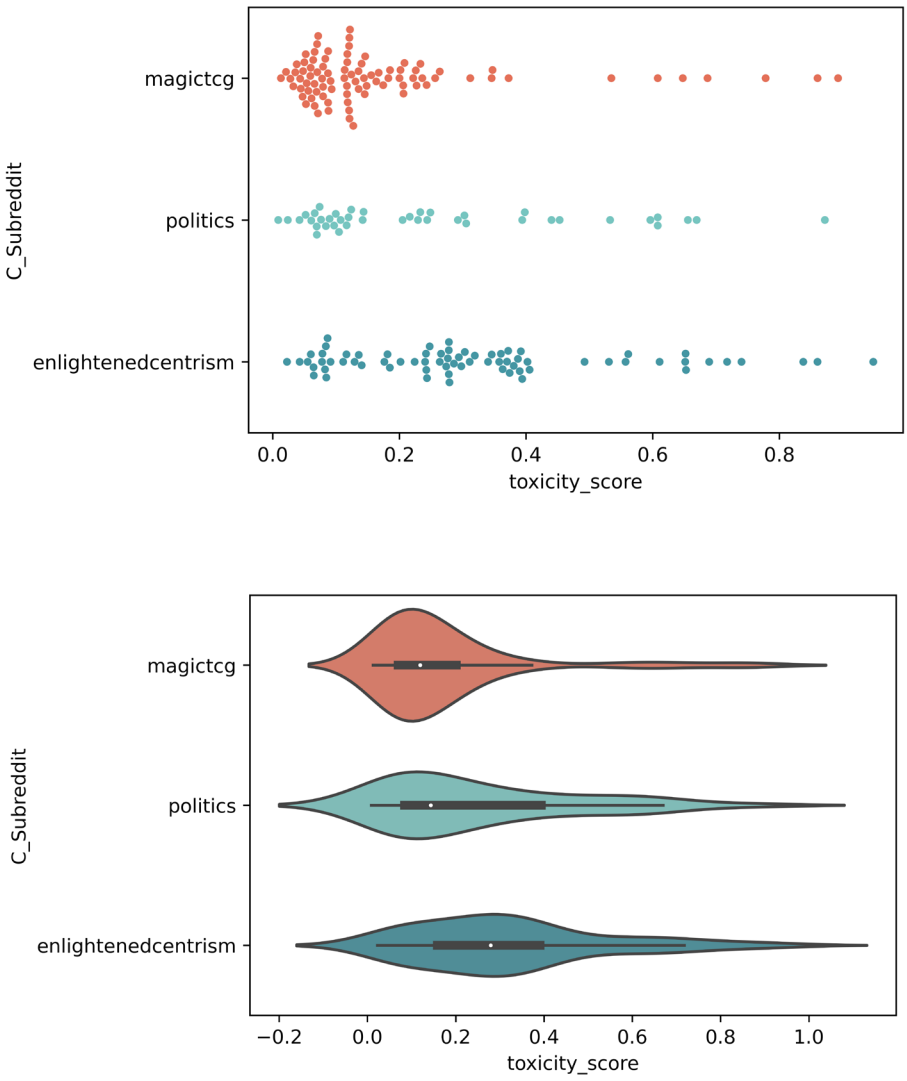


FIGURE 9.
COMMENTS LEFT BY A SINGLE USER ACROSS THREE SUBREDDITS, BY TOXICITY

These charts show clear differences in behaviour as this user moves between subreddits. /r/politics and /r/enlightenedcentrism are both spaces for political discussion, but this participant is less likely to be toxic in the former. This reflects the character of the subreddits - across all users, /r/enlightenedcentrism has the higher average comment toxicity. In /r/magictcg, one of the least toxic subreddits in our study, this change in behaviour is even more pronounced. If these charts show a person adapting to their environment, our aim here is to work out what is different about the environments which encourages these shifts.

To measure this effect at scale, we calculated a single 'adjustment' score for each subreddit, designed to indicate the extent to which users posting there change their tone. This is an average of the difference, for each user posting in a subreddit, between that user's comment toxicity elsewhere compared to that subreddit. A positive adjustment score indicates that people are likely to be more toxic in that space than they are elsewhere - a negative score indicates the opposite.

COMPARATIVE ANALYSIS

A series of linear regression models were built to investigate the relationship between each of the subreddit characteristics as independent variables - numbers of users, moderators and rule categories - and the toxicity measures as dependent variables - the size of Q3, the adjustment score and the percentage of comments over 0.8 toxicity.

Finding Three: Larger spaces contain a higher proportion of highly toxic comments.

Across each of our measures of toxicity, subreddits which contained a larger number of users tended to be less healthy - people became more toxic when commenting there, the provocateurs were prepared to be more provocative, and a higher proportion of comments were highly toxic. This last connection was the strongest, with the model predicting that, for every 1000 users in a space, another 1% of that space's comments will be highly toxic. This relationship is interesting as it conflicts with an intuition borrowed from offline space, that people are more likely to temper their behaviour in well populated, more public spaces.

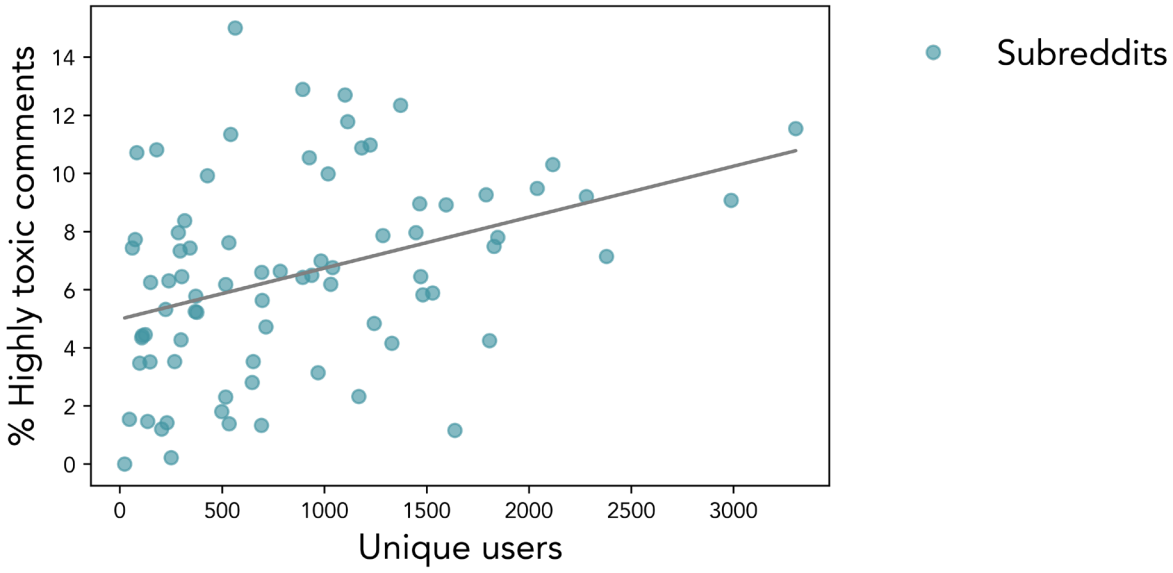


FIGURE 10.
NUMBER OF UNIQUE USERS COMMENTING ON A SUBREDDIT AGAINST PROPORTION OF HIGHLY TOXIC COMMENTS

As seen above, this isn't a direct relationship: the points on the scatter graph above don't conform too closely to the model's best fit line. The 'coefficient of determination' for this model, a measure of this distance, is 0.14. This suggests that the number of users in a space explains about 14% of the variance in toxicity between subreddits. This effect, then, is weak, though relatively high given the effect we might expect other variables, such as the subjects discussed on a subreddit - to have on this measure. Additionally, in the case studies below, some subjects' most toxic forums were the smallest. This measure, then, requires more careful attention.

Finding Four: Broader rule sets are associated with positive adjustments in toxicity.

The factor which had the most direct effect on a subreddit's heath was the number of rule categories present on that subreddit. In particular, broader rules were related to the amount which people were willing to adjust their general tone when commenting; the more rules a space declared, the more users reduced their comment toxicity to match that space. This reduction is small - our model predicts that every additional rule category of a possible 13 is associated with another 1% drop of toxicity from users posting there - but likely to be significant in helping maintain the norms within a space. Again, there is some variance in the graph above, and the model is only a loose fit, suggesting that rules account for 15% of the adjustment in toxicity.

This correlation does not explain how rules might cause this change. In particular, our data doesn't tell us whether the reduction in toxicity is a result of users reading the rules and tempering their behaviour appropriately, or whether rule-breaking comments were being made within a space and removed by attentive moderators. This result is likely to be a mixture of both. Either way, this relationship shows that this form of governance online can be effective -, whether they're being followed by users or moderators, the rules which Reddit allows its communities to set up have a measurable effect on how people behave in that space. We may therefore look to platforms to invest in their community moderation policies, rather than a narrow reliance on top-down systems.

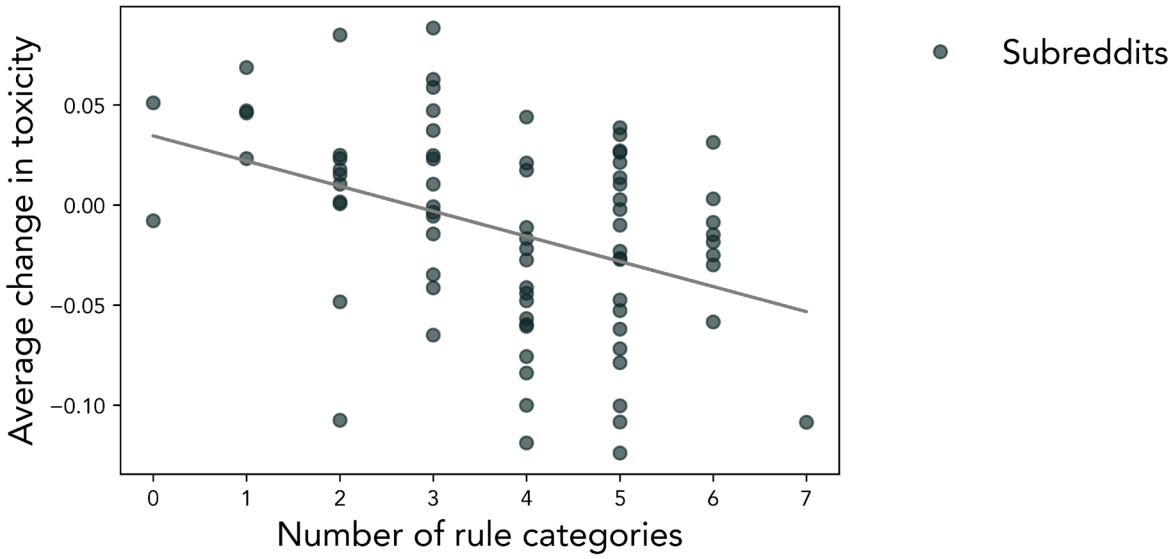


FIGURE 11.
NUMBER OF RULES IN A SUBREDDIT AGAINST THE AVERAGE USER CHANGE IN TOXICITY FOR THAT SPACE

Finding Five: Some individual rules appear to have a stronger impact on the average toxicity of a space than others.

In addition to the raw number of articulated rules likely impacting the relative toxicity of a space, there is evidence to suggest that the nature of those rules further impacts the toxicity of the forum and the discussions taking place therein. Analysts compared the average toxicity and the proportion of high toxicity comments by whether or not that forum displayed one or more rules in the categories described above. The results are shown in the table below.

Subreddit Rules	# Subreddits	Average Toxicity	% Comments 0.8 toxicity or higher
All subreddits	78	0.23	7.2%
Rules mandating civil behaviour	53	0.22	6.7%
Rules prohibiting bigotry, racism and hate	50	0.21	6.4%
Rules determining the structure of posts	44	0.23	7.3%
Rules prohibiting spam	43	0.23	6.8%
Posts & Comments must be on topic	41	0.22	6.7%
Rules around post and comment quality	41	0.22	6.6%
Rules against sharing of personal information	33	0.24	7.2%
No Memes	25	0.21	6.0%
No Politics	12	0.18	4.8%
Rules prohibiting sales and marketing	16	0.2	6.0%
Rules enshrining moderator discretion	10	0.23	8.3%
Rules banning certain user groups	8	0.26	9.3%
Rules restricting users' posting rights	6	0.23	7.2%

FIGURE 12.
TOXICITY MEASURES ACROSS SUBREDDITS CONTAINING DIFFERENT CATEGORIES OF RULES

Overall individual rules had a reasonably small effect on the toxicity when viewed across all subreddits, though some variety is noticeable. A ban on political discussion, for instance, appears to have had some impact in lowering both the average toxicity of a space and the number of highly toxic comments, while spaces explicitly banning participation by some group tend to contribute to a higher level of toxicity. This has further ramifications for how we develop political spaces online, if current political debate in spaces online contributes to a rise in toxicity.

Finding Six: The number of moderators had no effect on our toxicity measures

The sheer number of moderators in any subreddit was found to have no correlation to any of our measures of health. This is surprising - we might expect a higher number of eyes on a community to be helpful in building a set of norms, or contain fewer highly toxic comments. If moderation is having an effect on these measures of health, it is the quality, not the quantity, of moderators which is likely to be important.

Finding Seven: The number of comments made by a user has absolutely no effect on the average toxicity of the user's comments

Focusing on users, analysts tested the extent to which a user's activity correlated with the toxicity of their comments to investigate the hypothesis that a more active user may be more bought in or tuned in to the culture of a space. However, in line with the findings above that spaces may develop more or less toxic norms, we can show no increase or decrease in toxicity based on activity: an active user in a toxic environment is as likely to be toxic as a more infrequent user. This may suggest that, to improve behaviour, platforms should invest in building incentives for users to get involved in moderation of a range of online spaces they engage in over incentivizing users to engage more in a particular space where they are already active. For instance, Facebook could prioritise notifications about changes in moderation rules or admins to groups you are a member of above notifications that there is a new post in a group you are a member of on Facebook.

Finding Eight: A user's language does change in toxicity across different online spaces, but the effect is only extreme in a minority of cases

Analysts identified the 3,295 users who had posted at least five comments in two subreddits, then compared the average toxicity of comments across those subreddits to identify the difference

# Comments made	Average Toxicity
2-4	0.23
5-10	0.23
11-20	0.23
21-50	0.23
51-100	0.22
101+	0.23

FIGURE 13.
AVERAGE TOXICITY OF USERS BY NUMBER OF COMMENTS MADE

in average toxicity. The average toxicity of the subreddit was not taken into account at this stage, as the object is to identify whether a user changes their behaviour without judging the relative toxicity of the spaces they tend to occupy.

At one end of the spectrum were users who did not show any evidence of changing their language depending on the spaces they were posting in: of the 3,295 users analysed, 807 (24%) maintained consistent levels of toxicity. At the other end of the spectrum, 27 users (0.8%) saw a five or six band differential. The results are shown in the graph below.

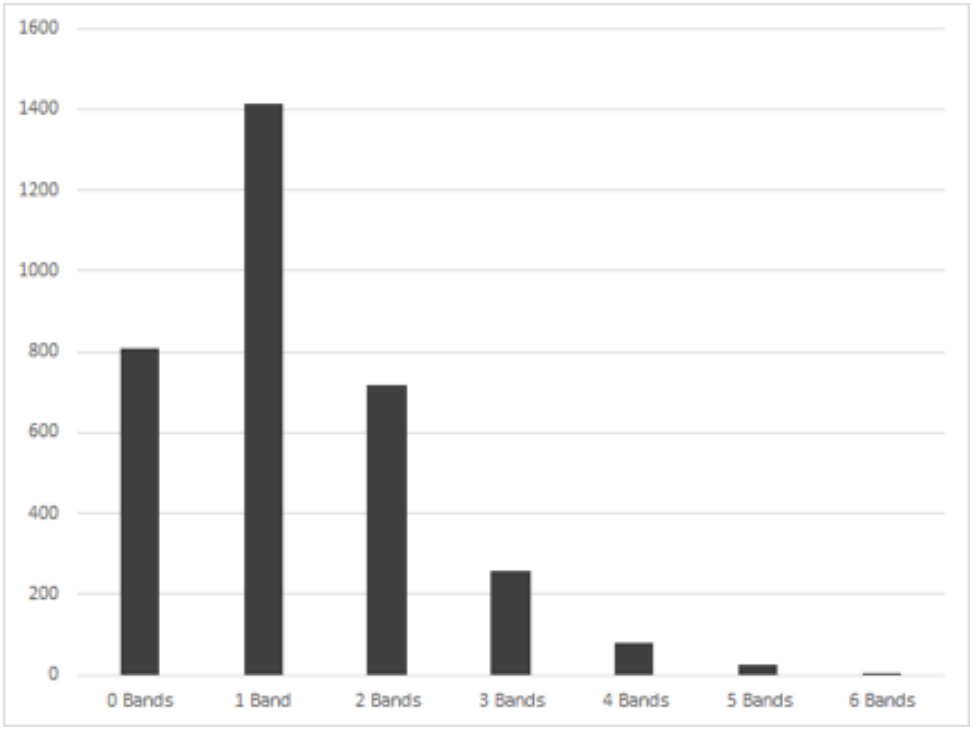


FIGURE 14.
DEGREES OF CHANGE IN SAME-USER LANGUAGE TOXICITY ACROSS ONLINE SPACES

The results show that the majority of the users (75%) in our dataset made at least some adjustment to their language depending on the spaces they were speaking in. Most often this was a small change: but 10% changed their language by three bands of toxicity or higher.

By way of illustration, one user's relative toxicity is shown below.

Author	Subreddit	# Comments	Average Toxicity
sd***	BreadTube	3	0.07
sd***	politics	5	0.70
sd***	science	11	0.12
sd***	worldnews	16	0.24

FIGURE 15.
AN EXAMPLE OF A USER'S RELATIVE TOXICITY ACROSS SUBREDDITS

When posting in the /r/worldnews, /r/science and /r/breadtube subreddits, the user's comments are classified as low toxicity, for example:

> *And even then, we're finding more and more that *any* amount of daily alcohol consumption isn't healthy (notwithstanding some low risk factors).*
>
> (Toxicity score: 0.0)

This changes when the user moves into /r/politics.

> *I personally do not give a fuck what people are into sexually... so fuck C** nearly as much as fuck Trump. That fuckers about to make a load of money by being a lying piece of shit in the book. Can't we get dirt without pouring money into these fuckers?*
>
> (Toxicity score: 0.8)

This differentiation in behaviour by groups of users across multiple spaces are the subject of three case studies below.

CASE STUDIES

The analysis above suggests there is a connection between the culture, norms and design of a space and the type of behaviour found therein. It also highlights the fact that a user, or a group of users, may act civilly in one space and less civilly in another.

The second question analysts looked to further explore was therefore the extent to which a space's toxicity was a product of a group of toxic users, or whether users moderated their behaviour depending on the conditions of the space as illustrated above. In short, are users consistently toxic across all spaces, and the solution is removing the bad eggs? Or is the problem the conditions a space creates for behaviour, and an otherwise civil user will behave badly in a more toxic environment?

To explore this question, we carried out three case studies examining a handful of subreddits which dealt - or were dealing with at the time of data collection - a single theme or topic. The ambition was to isolate users who were active across at least two of these to see how far their toxicity changed.

MAGIC: THE GATHERING

Magic: The Gathering is a trading-card game played by millions of people around the world, both in paper and online. Surrounding the game is a wide ecosystem of discussions forums, news reportage and gameplay content, including on Reddit. A handful of subreddits dedicated to the game were selected, aiming to capture a range of size, attitude and community standards. These subreddits, their size and descriptions, and their average comment toxicity are shown in the table below.

Subreddit	Description	Subscribers	Average Comment Toxicity
magicTCG	General Discussion Forum	422,000	0.11
MagicArena	General Discussion Forum	192,000	0.11
spikes	Competitive Discussion Forum	86,300	0.10
mtgfinance	Discussion of Magic: The Gathering finance	72,000	0.13
magicthecirclejerking	Oppositional Discussion Forum	55,000	0.18
freemagic	Oppositional Discussion Forum	9,800	0.28

FIGURE 16.
AVERAGE TOXICITY AND USER BASE PER SUBREDDIT

Oppositional discussion forums exist to facilitate conversations not permitted in the larger subreddits, or to host users who have been banned from participating in the larger forums. It should therefore not be surprising that the average toxicity in these spaces is significantly higher, as in the case of /r/freemagic, a space describing itself as "a wide open and mostly unmoderated subreddit to talk about Magic: The Gathering." This supports the proposition that where multiple similar online spaces with different cultures and moderation practices exist, users will engage in behaviour that is more toxic in the spaces where that toxicity is established, accepted or encouraged.

We collected 49,500 comments across the *Magic: The Gathering* subreddits. Users were filtered to find those users who had posted at least five times over multiple subreddits, leaving a dataset of 382 users.

To test the central question of whether individual users acted differently in different spaces, each user's comments in each space they had participated in were scored for toxicity, then averaged. For instance, one user, shown in the table below, was active in both the main subreddit for *Magic: The Gathering* and the oppositional FreeMagic subreddit.

User: Zx***	No. of comments	Average Toxicity of Comments
magicTCG	24	0.07
freemagic	228	0.41

FIGURE 17.
COMMENTS AND TOXICITY FOR A SINGLE USER

When posting in the main subreddit, the user's comments were broadly in line with the expected toxicity: in fact, they were slightly less toxic than the average content and well below any threshold of problematic behaviour. In the main subreddit, the user contributed to discussions around strategy and tactics and even offered an introduction to the game to a newer player, while in /r/freemagic they engaged in bitter arguments with a high degree of toxicity.

Another user, active in /r/freemagic and /r/MtgFinance displayed a similar pattern of behaviour.

User: Ni***	No. of comments	Average Toxicity of Comments
MtgFinance	8	0.15
freemagic	32	0.45

FIGURE 18.
COMMENTS AND TOXICITY FOR A SINGLE USER

In /r/MtgFinance, the user is cordial, polite, and engaged in discussions. One post is shown below.

> *Sorry for all asking but it's good to know more. how does it look? do you have a picture? the rebacked cards i have seen look pretty messy with fake wear to hide the job.*
>
> *I'd like to know what it looked like at first, to be worth rebacking to make money compared to when it was real and damaged like it was before. It's a strange card to see reebacked.*

In /r/freemagic, the user is far more aggressive.

> *How often do you need to reply to me with dumb shit? You don't know what you're talking about... I don't need to explain fucking anything to shit behaviour people like you. Check your shite attitude dickhead.*
>
> *Oh i'm gonna do what some asshole who can't post a comment without talking bullshit. You being ignorant doesn't mean i'm an idiot. Jesus christ you're stupid*

These examples hint that a user on the platform changes the toxicity of their comments based on the perceived rules and cultures of the different communities on that platform. Analysts grouped users based on the communities they shared, then reported the average toxicity of the comments left by each group across the two spaces, as well as the difference between the two spaces. The results are shown below. For instance, comments shared by users posting in both /r/MagicArena and/r/freemagic were by some degree more toxic in /r/freemagic than those they posted in /r/MagicArena.

Subreddits	freemagic	MagicArena	magicTCG	magicthecirclejerking	mtgfinance	spikes	Delta
MagicArena/freemagic	0.26	0.09					0.17
spikes/freemagic	0.24					0.10	0.14
magicTCG/freemagic	0.23		0.13				0.10
mtgfinance/freemagic	0.25				0.15		0.10
mtgfinance/MagicArena		0.19			0.11		0.08
magicthecirclejerking/MagicArena		0.13		0.20			0.07
magicthecirclejerking/magicTCG			0.10	0.17			0.07
spikes/magicthecirclejerking				0.17		0.11	0.06
mtgfinance/magicthecirclejerking				0.20	0.14		0.06
magicthecirclejerking/freemagic	0.23			0.18			0.05
spikes/mtgfinance					0.15	0.11	0.04
spikes/macicTCG			0.14			0.11	0.03
spikes/MagicArena		0.15				0.14	0.1
magicTCG/MagicArena		0.11	0.11				0.00
mtgfinance/magicTCG			0.15		0.15		0.00

FIGURE 19.
COMPARISONS OF USER BEHAVIOUR ACROSS TWO SUBREDDITS (CASE STUDY SUBSECTION)

By comparison, users posting in the more heavily moderated, generalist discussion forums (such as /r/MagicTCG and /r/MagicArena) maintained a consistent level of toxicity. This is some evidence that it is not simply that bad spaces online attract bad actors, or that toxic online communities are full of toxic users, but that users adapt to the norms of a community and their behaviour is correspondingly more or less toxic.

Covid-19

The Covid-19 pandemic has led to the creation of dozens of related discussion fora on Reddit. These range from news sharing spaces, to support groups, to medical discussions and anti-lockdown groups. Across these groups, there is a spectrum of experience and opinion, with spaces that encourage the continuation of measures against the virus and support the publicly-accepted evidence, while other spaces reject the evidence and the subsequent measures implemented by governments around the world.

We collected 78,400 comments across ten Covid-19 related subreddits. The size and average comment toxicity of each subreddit is shown below.

Subreddit	Description	Subscribers	Average Comment Toxicity
COVID19positive	General Discussion Forum for users diagnosed with Covid-19.	92,000	0.10
COVID19_support	Forum providing support for those affected by the pandemic.	30,000	0.12
China_Flu	General Covid-19 discussion forum created before the disease was named Covid-19.	109,000	0.17
collapse	Non-Covid-19 specialist forum focusing on the collapse of civilisation.	237,000	0.19
LockdownSkepticism	Discussion forum skeptical of government lockdown policies responding to Covid-19.	22,000	0.19
Coronavirus	General Discussion Forum	2,300,000	0.2
NoNewNormal	Forum for action and rejection against lockdown and health measures.	10,500	0.22
CoronavirusFOS	General Discussion Forum focusing on freedom of speech.	72,000	0.23
Trumpvirus	Anti-Trump forum focusing on the US response to Covid-19.	36,000	0.27
CoronavirusCirclejerk	Discussion and memes satirising established Covid-19 discussions and positions.	6,500	0.27

FIGURE 20.
AVERAGE TOXICITY AND USER BASE PER SUBREDDIT

Again, those users active in at least two of these forums were selected and their average toxicity across the two spaces was compared. An example of one user active in two subreddits is shown below.

User: Pg***	No. of comments	Average Toxicity of Comments
collapse	11	0.11
Trumpvirus	35	0.42

FIGURE 21.
COMMENTS AND TOXICITY FOR A SINGLE USER

Writing in the general /r/collapse subreddit, the user's comments are largely advice on how to prepare for infrastructural collapse, such as preserving food or moving off grid.

Buy some hectares and build a house. You can make a little one for under $15k if you do the labour yourself. You can get all of the plans and materials for $30k

In /r/Trumpvirus, the user's posts are significantly more aggressive.

> *We loathe him; we loathe him. Trust me. I have fantasies about strangling his tiny pencil neck. I wake up in the middle of the night sometimes hating the asshole so much I can't get back to sleep. If there really was a devil it would be this sorry twat. I HATE HIM!*

Analysts grouped users based on the communities they shared, then reported the average toxicity of the comments left by each group across the two spaces, as well as the difference between the two spaces. Selected results are shown below: the ten subreddits with the highest deltas in user toxicity are shown.

Subreddits	China_Flu	Collapse	coronavirus	CoronavirusCirclejerk	CoronavirusFOS	COVID19_support	COVID19positive	LockdownSkepticism	NoNewNormal	Trumpvirus	Delta
China_Flu/CoronavirusCirclejerk	0.27			0.54							0.28
CoronavirusFOS/Coronavirus			0.34		0.16						0.18
CoronavirusCirclejerk/Coronavirus			0.13	0.29							0.16
Coronavirus/COVID19positive			0.22				0.07				0.15
Coronavirus/NoNewNormal			0.18						0.28		0.10
China_Flu/COVID_19support	0.19					0.09					0.09
CoronavirusCirclejerk/China_Flu	0.11			0.20							0.09
China_Flu/ COVID19positive	0.19						0.11				0.08
COVID_19support/LockdownSkepticism						0.17		0.24			0.07

FIGURE 22.
COMPARISONS OF USER BEHAVIOUR ACROSS TWO SUBREDDITS (CASE STUDY SUBSECTION)

Again, there is some divergence between users' toxicity in two spaces. Participation in the Covid-19 forums that promote freedom of speech and moderate less harshly led to users posting comments more frequently categorised as more toxic: the groups of users with the widest variation in their toxicity were active in at least one of these alternative, laxly moderated spaces. With ten subreddits analysed, however, the relative divergence was reasonably low - the average delta across each group was just 0.04. This suggests that unless a forum's culture and approach to moderation is meaningfully divergent from the main, user behaviour across multiple forums is for the most part stable. Nine of the user groups analysed showed no divergence at all. It is worth noting that this consistency was not limited to low-toxicity groups: users posting to /r/LockdownSkepticism and /r/NoNewNormal had an average toxicity of 0.21 in both cases.

US Policing and Police Violence

A third case study examined a handful of subreddits hosting discussions of policing and police violence, with data collection focusing on the dates and locations of the protests in the city of Portland in the US. Data was collected from a range of subreddits linked to discussions of law and order, and that host a range of perspectives on the roles, responsibilities and expectations of police in the US.

We collected far fewer comments here: 20,300 across the eight subreddits. Those subreddits are shown below.

Subreddit	Description	Subscribers	Average Comment Toxicity
askportland	Forum for questioning Portland residents.	21,000	0.11
TalesFromTheSquadCar	Forum for storytelling by members of law enforcement.	103,000	0.19
Good_Cop_Free_Donut	Discussion forum for positive and complementary discussions of policing.	35,000	0.20
Military	General military discussion forum.	280,000	0.22
ProtectAndServe	Law and order discussion forum; majority pro-police discussions.	186,000	0.22
Portland	General Discussion Forum for Portland, Oregon	179,000	0.23
BlackLivesMatter	News and Discussion related to the Black Lives Matter movement.	98,000	0.24
Bad_Cop_No_Donut	Forum centred on police brutality and abuses of power by law enforcement.	467,000	0.31

FIGURE 23.
AVERAGE TOXICITY AND USER BASE PER SUBREDDIT

Where subreddits represented factions in a conflicting debate, it was unsurprising to note increasing toxicity when participating in a forum with an alternative perspective. A user who posted regularly on /r/ProtectAndServe with an average toxicity of 0.14 posted a handful of times in /r/BlackLivesMatter with comments averaging 0.64 toxicity. Comments from the two subreddits are shown below, respectively.

> *Agreed definitely. If they respond to a crime they should politely ask the criminal to stop and if he refuses the officer should just move along*

> *It's ok to deface public property with black lives matter but it's not ok to do it with a blue lives matter mural. Why because it doesn't fit your narrative. Hypocrites*

Over the analysis period, only 144 users were active (defined as five or more comments left) in two subreddits, meaning that the findings for this case study can only be counted as indicative or otherwise of wider trends we have observed during this research. Nevertheless, analysts ran a similar analysis, identifying the difference in average toxicity by users across two subreddits.

Subreddits	askportland	Bad_Cop_No_Donut	BlackLivesMatter	Good_Cop_Free_Donut	Military	Portland	ProtectAndServe	TalesFromTheSquadCar	Delta
Bad_Cop_No_Donut/BlackLivesMatter		0.48	0.17						0.31
Bad_Cop_No_Donut/ProtectAndServe		0.35					0.15		0.21
BlackLivesMatter/ProtectAndServe			0.42				0.25		0.17
Bad_Cop_No_Donut/Portland		0.37				0.23			0.14
TalesFromTheSquadCar/Military					0.23			0.12	0.11
Good_Cop_Free_Donut/Bad_Cop_No_Donut/		0.27		0.17					0.10
Bad_Cop_No_Donut/Military		0.37			0.28				0.09
Military/TalesFromTheSquadCar					0.20			0.11	0.09
Good_Cop_Free_Donut/ProtectAndServe				0.31			0.24		0.08
Good_Cop_Free_Donut/TalesFromTheSquadCar				0.23				0.16	0.07
ProtectAndServe/ TalesFromTheSquadCar							0.19	0.16	0.03
Portland/ TalesFromTheSquadCar						0.24		0.27	0.03
TalesFromTheSquadCar/ ProtectAndServe							0.20	0.17	0.03
Military/ProtectAndServe					0.22		0.25		0.03
Bad_Cop_No_Donut/TalesFromTheSquadCar		0.20						0.22	0.01
Portland/askportland	0.16					0.15			0.01
TalesFromTheSquadCar/ Good_Cop_Free_Donut				0.20				0.21	0.01

FIGURE 24.
COMPARISONS OF USER BEHAVIOUR ACROSS TWO SUBREDDITS (CASE STUDY SUBSECTION)

Again, there is a familiar pattern, insufficient data notwithstanding. Users adjust their toxicity depending on the spaces they are in. The same user posting in /r/Bad_Cop_No_Donut at an average of nearly 0.5 toxicity posts with far greater civility in /r/BlackLivesMatter. It is the norms, moderation practices or culture of a space that shapes the conversations in that forum, rather than the user's tendencies.

This case study complicates matters further than, say, the case study examining the *Magic: The Gathering* trading card game, by introducing forums where users potentially conflict with one another in worldview or opinion. Anecdotally, given the small amount of data gathered here, we do see that subreddits with differing positions on the issues facing the police and the policed see divergent levels of toxicity by the users who participate in both. Spaces with similar cultures or worldviews see relatively stable levels of toxicity across the two spaces. Users posting in /r/ProtectAndServe tend to be more civil there than when posting in /r/Bad_Cop_No_Donut. Conversely, users posting in both /r/TalesFromTheSquadCar and /r/Good_Cop_Free_Donut maintain similar levels of toxicity in both. Although we might expect that users may join adversarial communities to troll or argue with their members, this data is cannot show that more conclusively. It is, however, sufficient to show divergent behaviour by the same groups of users.

Finding Nine: Users in online spaces do not behave consistently across multiple spaces, but adjust their toxicity or civility to the norms, rules or cultures of the spaces they use

CONCLUSIONS

This study highlights a need for new perspectives in thinking about the proliferation of online harms that moves beyond user behaviour and takes into account the design, custody and cultures of individual online communities. There are good spaces online, and there are bad spaces online, and the same groups of users frequently inhabit both. This dissonance can be found on all major platforms: a user may be simultaneously participating in the Facebook group for a local community organisation, their office Slack, far-Right Twitter conversations and dedicated hobby or interest sites. Each of these spaces is different, and this analysis suggests that a user will behave differently in each one.

We show that measuring the relative health of a space is possible. In fact, with open source algorithms, it can be done efficiently and cheaply, and further development of this methodology should allow for new ways to characterize how healthy or unhealthy a space has become. The methodology might take into account the size or rules or moderation practices, as this study has attempted, or may branch into alternative, computational measures. It is a cliché to call for further research, but in this instance it feels appropriate to.

This research highlights a key challenge in the realisation of online platforms as the new 'public forum' where people can come together to discuss and have democratic debates: namely, that political spaces tend to be more toxic than other spaces: and that users who are otherwise helpful and generous are more likely in a political space to become rude and aggressive. This is not a problem that post-hoc content moderation can solve alone: shaping user behaviour as they enter a space will be more productive in reducing toxic discussion than trying to stamp out bad behaviour after it has arisen. If the aim is to encourage a better quality of discussion online, policies to support the establishment of clear social norms for individual communities within these platforms are likely to show more returns than only targeting individuals who may seem to be 'toxic'. Platforms should thus invest in their moderators: and simply hiring more platform moderators may not be as effective as channelling those resources to upskill, empower and incentivise high-quality community moderation.

Significantly, we show that the narrow focus of much commentary on 'bad eggs' online is insufficient. Of course, an individual or group of users can influence an online space, and others through it: the online radicalisation of extremists is a useful reminder of this. But it is also important to take into account the ways in which an online platform may make that process more or less easy, the breadth of different platforms a user might be participating in, and what steps can be taken by regulators, platform architects and community members to shape spaces where toxicity and antisocial behaviour is reduced for all participants. This should be a source of optimism: that it is not simply we have to learn to defend the internet from inevitable bad actors, but that good design of online spaces itself can promote healthier behaviour and interactions - preventing, rather than removing, toxic discussion.

License to publish

Demos – License to Publish

The work (as defined below) is provided under the terms of this license ('license'). The work is protected by copyright and/or other applicable law. Any use of the work other than as authorized under this license is prohibited. By exercising any rights to the work provided here, you accept and agree to be bound by the terms of this license. Demos grants you the rights contained here in consideration of your acceptance of such terms and conditions.

1 Definitions

a 'Collective Work' means a work, such as a periodical issue, anthology or encyclopedia, in which the Work in its entirety in unmodified form, along with a number of other contributions, constituting separate and independent works in themselves, are assembled into a collective whole. A work that constitutes a Collective Work will not be considered a Derivative Work (as defined below) for the purposes of this License.

b 'Derivative Work' means a work based upon the Work or upon the Work and other pre-existing works, such as a musical arrangement, dramatization, fictionalization, motion picture version, sound recording, art reproduction, abridgment, condensation, or any other form in which the Work may be recast, transformed, or adapted, except that a work that constitutes a Collective Work or a translation from English into another language will not be considered a Derivative Work for the purpose of this License.

c 'Licensor' means the individual or entity that offers the Work under the terms of this License.

d 'Original Author' means the individual or entity who created the Work.

e 'Work' means the copyrightable work of authorship offered under the terms of this License.

f 'You' means an individual or entity exercising rights under this License who has not previously violated the terms of this License with respect to the Work, or who has received express permission from Demos to exercise rights under this License despite a previous violation.

2 Fair Use Rights

Nothing in this license is intended to reduce, limit, or restrict any rights arising from fair use, first sale or other limitations on the exclusive rights of the copyright owner under copyright law or other applicable laws.

3 License Grant

Subject to the terms and conditions of this License, Licensor hereby grants You a worldwide, royalty-free, non-exclusive, perpetual (for the duration of the applicable copyright) license to exercise the rights in the Work as stated below:

a to reproduce the Work, to incorporate the Work into one or more Collective Works, and to reproduce the Work as incorporated in the Collective Works;

b to distribute copies or phono-records of, display publicly, perform publicly, and perform publicly by means of a digital audio transmission the Work including as incorporated in Collective Works; The above rights may be exercised in all media and formats whether now known or hereafter devised. The above rights include the right to make such modifications as are technically necessary to exercise the rights in other media and formats. All rights not expressly granted by Licensor are hereby reserved.

4 Restrictions

The license granted in Section 3 above is expressly made subject to and limited by the following restrictions:

a You may distribute, publicly display, publicly perform, or publicly digitally perform the Work only under the terms of this License, and You must include a copy of, or the Uniform Resource Identifier for, this License with every copy or phono-record of the Work You distribute, publicly display, publicly perform, or publicly digitally perform. You may not offer or impose any terms on the Work that alter or restrict the terms of this License or the recipients' exercise of the rights granted hereunder. You may not sublicense the Work. You must keep intact all notices that refer to this License and to the disclaimer of warranties. You may not distribute, publicly display, publicly perform, or publicly digitally perform the Work with any technological measures that control access or use of the Work in a manner inconsistent with the terms of this License Agreement. The above applies to the Work as incorporated in a Collective Work, but this does not require the Collective Work apart from the Work itself to be made subject to the terms of this License. If You create a Collective Work, upon notice from any Licensor You must, to the extent practicable, remove from the Collective Work any reference to such Licensor or the Original Author, as requested.

b You may not exercise any of the rights granted to You in Section 3 above in any manner that is primarily intended for or directed toward commercial advantage or private monetary compensation. The exchange of the Work for other copyrighted works by means of digital file sharing or otherwise shall not be considered to be intended for or directed toward commercial advantage or private monetary compensation, provided there is no payment of any monetary compensation in connection with the exchange of copyrighted works.
c If you distribute, publicly display, publicly perform, or publicly digitally perform the Work or any Collective Works, you must keep intact all copyright notices for the Work and give the Original Author credit reasonable to the medium or means You are utilizing by conveying the name (or pseudonym if applicable) of the Original Author if supplied; the title of the Work if supplied. Such credit may be implemented in any reasonable manner; provided, however, that in the case of a Collective Work, at a minimum such credit will appear where any other comparable authorship credit appears and in a manner at least as prominent as such other comparable authorship credit.

5 Representations, Warranties and Disclaimer
a By offering the Work for public release under this License, Licensor represents and warrants that, to the best of Licensor's knowledge after reasonable inquiry:
i Licensor has secured all rights in the Work necessary to grant the licence rights hereunder
and to permit the lawful exercise of the rights granted hereunder without You having any obligation to pay any royalties, compulsory licence fees, residuals or any other payments;
ii The Work does not infringe the copyright, trademark, publicity rights, common law rights or any other right of any third party or constitute defamation, invasion of privacy or other tortious injury to any third party.
b Except as expressly stated in this licence or otherwise agreed in writing or required by applicable law, the work is licenced on an 'as is' basis, without warranties of any kind, either express or implied including, without limitation, any warranties regarding the contents or accuracy of the work.

6 Limitation on Liability
Except to the extent required by applicable law, and except for damages arising from liability to a third party resulting from breach of the warranties in section 5, in no event will licensor be liable to you on any legal theory for any special, incidental, consequential, punitive or exemplary damages arising out of this licence or the use of the work, even if licensor has been advised of the possibility of such damages.

7 Termination
a This License and the rights granted hereunder will terminate automatically upon any breach by You of the terms of this License. Individuals or entities who have received Collective Works from You under this License, however, will not have their licences terminated provided such individuals or entities remain in full compliance with those licences. Sections 1, 2, 5, 6, 7, and 8 will survive any termination of this License.
b Subject to the above terms and conditions, the licence granted here is perpetual (for the duration of the applicable copyright in the Work). Notwithstanding the above, Licensor reserves the right to release the Work under different licence terms or to stop distributing the Work at any time; provided, however that any such election will not serve to withdraw this License (or any other licence that has been, or is required to be, granted under the terms of this License), and this License will continue in full force and effect unless terminated as stated above.

8 Miscellaneous
a Each time You distribute or publicly digitally perform the Work or a Collective Work, Demos offers to the recipient a licence to the Work on the same terms and conditions as the licence granted to You under this License.
b If any provision of this License is invalid or unenforceable under applicable law, it shall not affect the validity or enforceability of the remainder of the terms of this License, and without further action by the parties to this agreement, such provision shall be reformed to the minimum extent necessary to make such provision valid and enforceable.
c No term or provision of this License shall be deemed waived and no breach consented to unless such waiver or consent shall be in writing and signed by the party to be charged with such waiver or consent.
d This License constitutes the entire agreement between the parties with respect to the Work licensed here. There are no understandings, agreements or representations with respect to the Work not specified here. Licensor shall not be bound by any additional provisions that may appear in any communication from You. This License may not be modified without the mutual written agreement of Demos and You.

DEMOS

Demos is a champion of people, ideas and democracy. We bring people together. We bridge divides. We listen and we understand. We are practical about the problems we face, but endlessly optimistic and ambitious about our capacity, together, to overcome them.

At a crossroads in Britain's history, we need ideas for renewal, reconnection and the restoration of hope. Challenges from populism to climate change remain unsolved, and a technological revolution dawns, but the centre of politics has been intellectually paralysed. Demos will change that. We can counter the impossible promises of the political extremes, and challenge despair – by bringing to life an aspirational narrative about the future of Britain that is rooted in the hopes and ambitions of people from across our country.

Demos is an independent, educational charity, registered in England and Wales. (Charity Registration no. 1042046)

Find out more at **www.demos.co.uk**

DEMOS

PUBLISHED BY DEMOS MARCH 2021.

15 WHITEHALL, LONDON, SW1A 2DD
T: 020 3878 3955
HELLO@DEMOS.CO.UK
WWW.DEMOS.CO.UK