

DEMOS

EVERYTHING IN MODERATION

PLATFORMS, COMMUNITIES
AND USERS IN A HEALTHY
ONLINE ENVIRONMENT

ALEX KRASODOMSKI-JONES

OCTOBER 2020

Open Access. Some rights reserved.

Open Access. Some rights reserved. As the publisher of this work, Demos wants to encourage the circulation of our work as widely as possible while retaining the copyright. We therefore have an open access policy which enables anyone to access our content online without charge. Anyone can download, save, perform or distribute this work in any format, including translation, without written permission. This is subject to the terms of the Creative Commons By Share Alike licence. The main conditions are:

- Demos and the author(s) are credited including our web address **www.demos.co.uk**
- If you use our work, you share the results under a similar licence

A full copy of the licence can be found at **<https://creativecommons.org/licenses/by-sa/3.0/legalcode>**

You are welcome to ask for permission to use this work for purposes other than those covered by the licence. Demos gratefully acknowledges the work of Creative Commons in inspiring our approach to copyright. To find out more go to **www.creativecommons.org**



This project was supported by GCHQ



Published by Demos October 2020

© Demos. Some rights reserved.

15 Whitehall, London, SW1A 2DD

T: 020 3878 3955

hello@demos.co.uk

www.demos.co.uk

CONTENTS

ACKNOWLEDGEMENTS	PAGE 4
INTRODUCTION	PAGE 5
ONLINE HARMS LEGAL vs ILLEGAL	PAGE 8
PART 1 PLATFORM LEVEL MODERATION AND THE TERMS OF SERVICE	PAGE 9
PART 2 COMMUNITY MODERATION	PAGE 14
PART 3 SELF-MODERATION AND ONLINE CULTURES AND NORMS	PAGE 22
PART 4 RECOMMENDATIONS	PAGE 25
RECOMMENDED READING	PAGE 26

ACKNOWLEDGEMENTS

The generosity shown by the contributors to this paper in giving up their time to explore how constitutional law, criminology, social control and policing may shed light on the challenges of the online world was remarkable. My thanks to Martin Innes, Ian Loader, David Leeney, Rick Muir, Giovanni De Gregorio and other interviewees. Thanks also to the team at GCHQ for supporting and informing this work, the team at Demos for their insights and careful review, and to Joe Cooke.

Alex Krasodonski-Jones

September 2020

ABOUT THE GOOD WEB PROJECT

The Good Web Project will articulate the vision for an internet compatible with liberal democracy. We will create the evidence base and principles for policy makers and opinion leaders in liberal democracies worldwide to advocate for an internet in robust contrast to authoritarian models. We believe that with a mix of policy, dialogue and a granular understanding of technology, the internet will be the place where democracy is redefined in the 21st century.

INTRODUCTION

It is barely an oversimplification to characterise current debate on internet regulation as a fight over the things people see, and the things they don't. The systems of curation and moderation that dictate what is and isn't permitted are the machinery most responsible for the health of online spaces. It follows that the ways they work are the subject of intense government scrutiny.

This paper argues that the principle and practice underpinning most major platforms have failed to create healthy online spaces, and that current attempts by states to regulate these spaces will in all likelihood fall short of addressing the root causes of this failure.

In short, we identify three failures in current approaches to content moderation.

- There is a democratic deficit in the way the majority of online platforms are moderated both in principle and in practice.
- The architecture of the majority of online platforms undermine the abilities of communities to moderate themselves both in principle and in practice.
- The majority of online platforms lack the cultures and norms that in the offline world act as a bulwark against supposed harms.

The first is a shortcoming in governance. Major platforms have taken on a public and state-like role in societies, including the writing and enforcement of rules that govern acceptable behaviour, in the manner of the private companies that they are. The rules governing online spaces and the ways in which those rules are enforced are primarily a function of profit. Platform decision-making, processes and technologies are undemocratic, unaccountable, opaque and its users lack reasonable means of redress when things go wrong.¹ This approach to governance is authoritarian at worst and at best a subjugation of public values before commercial interest.

The second is a shortcoming in architecture. The systems used by the majority of major platforms to police themselves are narrow and top-down. They fail to empower or incentivise communities to meaningfully police themselves. No number of digital citizenship and digital education initiatives will be impactful while online spaces fail to provide the mechanisms to turn good intentions into action. By way of analogy, plural policing and pro-social control of offline society by groups and individuals - from religious groups to healthcare providers, family and education systems to private security - is poorly supported by current platform design.

The third sits somewhere across the two. Design failures inhibit citizen participation. A democratic deficit inhibits a sense of digital citizenship. A disincentivised, disengaged and powerless user base produces corresponding cultures and norms. Uniquely digital effects likely impact the shape of online communities still further: unstable identities, online disinhibition and the vast scale and heterogeneity of some online communities, for examples.

In summary, unaccountable, authoritarian regimes set and enforce profit-maximising rules on powerless populations that lack the means or incentives to build out the pro-social structures and cultures upon which a healthy, democratic society depends.

Improving these spaces therefore requires platforms to answer these three challenges.

The first challenge is moving from authoritarianism towards democracy. Insofar as platforms play a quasi-public role, their processes should be subject to public scrutiny. This requires a realignment, from acting solely in a private interest to also acting in a public interest. Public roles in a democracy require transparency, accountability and the right to redress, and are built to reflect the rights and values of their publics. The rules and processes of content moderation should be too. It must be made clear to a member of the community why behaviour is

1. See, for instance, T. Gillespie, *Custodians of the Internet. Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media* (Yale University Press 2018)

or isn't allowed, how that judgement is made, and where to appeal that judgement if it is seen to have been made in error.

The second challenge is to turn digital subjects into digital citizens. Limiting the powers of platforms by the fundamental rights of their users - digital constitutionalism - lays a foundation on which a platform must provide the tools, structures and incentives for users to actively participate in shaping society, digital or otherwise. Platforms should embrace technologies and processes that enable and reward civic labour that shapes a community in line with that community's laws, rules and terms of service. Certain platforms lead the way here through implementing financial incentives, reputational incentives and user-led systems of content moderation.

Finally, we require the development of cultures conducive to minimising online harm. This is challenging, as the Internet hosts communities of an immeasurable range of perspectives, values and norms. We propose two answers. First, that certain values are better than others. Promoting values of respect, understanding and equality, as well as fundamental human rights as put forward by the United Nations, should be encouraged. Second, that the infrastructure on which online communities and cultures are built should help empower and inform that community. Inevitably there will be parts of the web that host communities we disagree with. We believe that implementing infrastructural change, such as steps to ensure plurality of opinion and information, the empowerment of users and the stabilisation of identity will move the dial while avoiding accusations of censorship and authoritarian overreach.

Alongside changes to platforms, progress here will require states to rethink their approach to online regulation. For one, governments will have to stop jealously peering over the fence at the apparent successes authoritarian regimes have in controlling the digital commons. Switching the Internet off is a sure-fire way of preventing online harms, but hardly an approach that is consistent with liberal democratic principles.

States must also redefine how they understand success. We view the use of force as a failure in the offline world: a stop-gap necessity used to patch a tear in the social fabric. Force buys the time needed for social, economic and political forces

to heal the wound. Deploying the police is a sign that something has gone wrong, not a sign that something is working. Resorting to bans, blocks, takedowns and censorship online should be treated in the same way. Outside of a narrow range of illegal behaviour, measuring the effectiveness of a platform's stewardship by the speed and volume of forceful interventions is a mistake.

Instead, states should point to those parts of the web that act as beacons for what Good might look like. Fewer 8chans is a worthy goal, but a myopic one. It should be complemented by lobbying for more Wikipedias, more StackOverflows, more Bumbles and so on. We cannot focus solely on what we don't want, and forget about what we do.

This paper explores content moderation in three strata, shown in Fig. 1 below. First, we examine the democratic deficit in the way the majority of online platforms draw up and enforce their rules. Second, we explore the failures in platform design that undermine the abilities of communities to moderate themselves. Finally, we raise the need for developed cultures and norms that may act as a bulwark against supposed harms. We approach each area by presenting problems and possible solutions, both in principle and in practice. The paper concludes with recommendations.

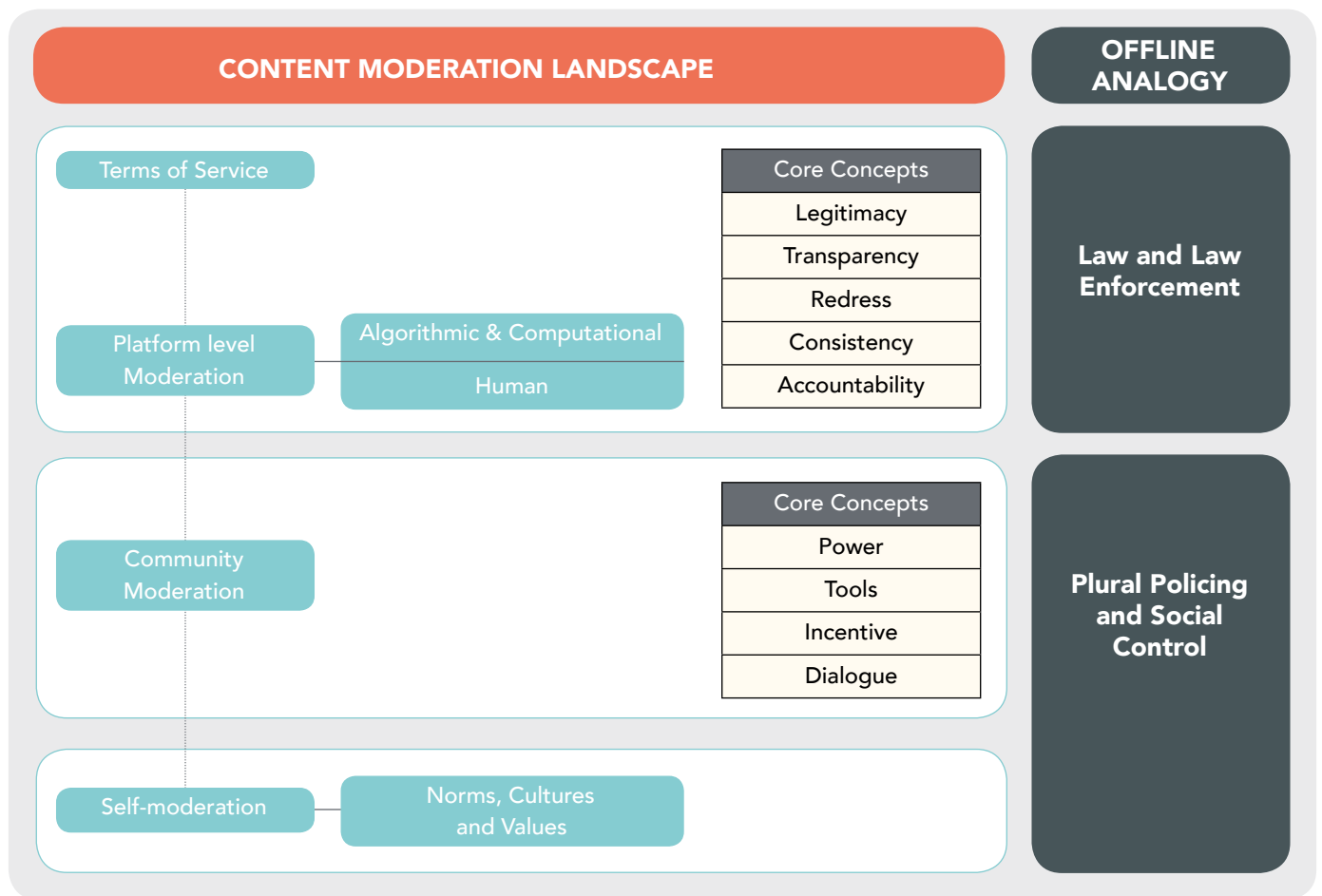


FIGURE 1.
THE LAYERS OF THE CONTENT
MODERATION LANDSCAPE

ONLINE HARMS LEGAL vs ILLEGAL

The focus of this paper is primarily on moderation as a tool of social order, and a force for improving the health of online spaces. It reflects ongoing UK discussions about online harms, focusing on those “harmful but legal” behaviours at the center of the policy debate. The paper explores how the principles of liberal democracy could be better written into the codebases underpinning the online spaces we increasingly inhabit, and how future regulation might most effectively encourage and secure a healthier web.

However, there are crimes committed online that fall squarely and urgently under the remit of law enforcement. Where this is the case, debates over citizenship, values and norms lose meaning. Internationally, states have come to a consensus on serious crime: exploitation of children, human trafficking and slavery are not so much questions of social order as they are problems that need policing solutions. In the debate about the policing mission, it is clear that these harms constitute problems where forceful intervention is justified and legitimate. Technological solutions to fight these crimes specifically, and the powers required to implement them, are beyond the scope of the paper, but where they can be shown to be effective should be implemented and enforced.

PART 1

PLATFORM LEVEL MODERATION AND THE TERMS OF SERVICE PLATFORM

THE PROBLEM IN PRINCIPLE

At its most basic, there is a democratic deficit in the way major online platforms are policed. We see shortfalls in legitimacy, accountability, transparency, and means of individual and collective redress. We have expectations as to the way rules are set and enforced in liberal democracies, and these expectations offer a useful framework for understanding the shortcomings of the platform model.

Under the platform model, acceptable behaviour is set and enforced by private actors. This presents a tension. On the one hand, platforms are compelled to protect users' freedom of expression as the core purpose of their product, and their policy rhetoric reflects this. On the other, content must be policed, moderated and curated in order for a platform to remain functional and attractive, that is to say, for profit. It is the protection of the commercial interests of online platforms that determines the rules by which its users must behave and the content that flows through the communities they inhabit, rather than, say, public values.

This shouldn't be surprising: private companies are subject to vastly different expectations than public bodies. Facebook is not a governmental actor.

Nevertheless, lawmakers around the world are recognising that online platforms have displaced streets and parks as "the most important places . . . for the exchange of views," as Justice Kennedy in the US has commented.² Mark Zuckerberg himself has talked publicly about Facebook's state-like pretenses.

*"In a lot of ways Facebook is more like a government than a traditional company... We have this large community of people, and more than other technology companies we're really setting policies."*³

As platforms take on ever-greater public and quasi-public roles, their compatibility with democracy and the legitimacy of their decisions will continue to be questioned.

The importance of injecting public values to democratise and legitimise security in policing is the subject of work by Loader and White. Writing in 2015, they lay out a requirement that regulation of private policing capabilities should aim to civilise the industry.⁴ That is to say, regulation "where the non-contractual public values and commitments of both market and non-market actors can be expressed, deliberated upon, and (if appropriate) institutionalized." Morality meets the market.

2. *Packingham v. North Carolina*, 137 S. Ct. 1730, 1737–38 (2017).

3. F. Foer, *Facebook's war on free will: How technology is making our minds redundant*, *Guardian* (Sept 2017).

4. I Loader & A. White, *How can we better align private security with the public interest? Towards a civilizing model of regulation* (Regulation and Governance 2015)

In *Revisiting the Police Mission*, Ian Loader writes:

*In a liberal democracy, it matters not simply that crime is controlled and order maintained. It also matters greatly how crime is controlled and what kind of order is maintained. Once one acknowledges the import of these wider considerations (as I think we must), we are necessarily drawn into a discussion about such matters as the power and limits of the state, the rights of the individual, the virtues and dangers of community, and the organisation of political authority.*⁵

We can draw useful parallels to the online world here. Our concerns should not stop at whether or not a platform is policed or made orderly, but rather the manner in which it is. In the dominant platform model, arbitrary and opaque terms of service drawn up to minimise economic risk are enforced through private law, moderation teams and algorithms. Moreover, criminologists note that when security is implemented and treated as an economic-driven necessary, or commodified in one way or another, effectiveness must be balanced against cost, and risks prioritising valuable assets over less valuable ones. This is reflected in platforms' heavy investment in English-speaking content moderation teams, for example. The experience of US users on the platform is ten times more valuable than the experiences of much of the global South when measured in advertising revenue.⁶

If top-down, search-and-remove policing is the way problems are solved, and the only way, all problems risk becoming viewed through that lens. We can see evidence of this myopia in much of the discussion around the health of online platforms, as terrorist content and child sexual abuse imagery (CSAI) is often bundled together with disinformation, offensive content and sexting as content that is harmful and ought to be policed away, in one way or another. Online harms present an enormous challenge to the mission and practice of policing, but the point made here is that the policing lens itself is at best insufficient and at worst a flat-out mistake.

Finally, users have no right nor route to contest the decisions made by higher powers under the default platform model. "Within this framework," writes Giovanni De Gregorio, "the lack of any users' rights or remedy leads online platforms to

exercise the same discretion of an absolute power over its community."⁷ Shoshana Zuboff calls these "the social relations of a pre-modern absolutist authority".⁸ Others have called the platform model feudalistic or Hobbesian: a system under which you give up your rights in exchange for products and services.⁹ Whatever it is, the current situation does not sit comfortably with our conception of citizens in a democracy. "The status quo system," writes Karl Langvardt, "in which private companies have free rein to design censorship protocols beyond the rule of law are almost shockingly dystopian when considered from a distance."¹⁰

THE PROBLEM IN PRACTICE

Failures in principle manifest in failures in practice. It is at its most stark in the ongoing debate about platform censorship. A day doesn't pass without one high-profile provocateur or another being removed from a platform, or not. The latest, Katie Hopkins here in the UK, had walked this tightrope for years before Twitter finally suspended her presence on the platform in June 2020 on the grounds her speech violated their terms and conditions.

Yet it remains unclear why her latest tweets were deemed enough, after years of provocative and offensive material. Perhaps the public outcry was simply too loud to be ignored, this time, but it is not clear why the platform made the decision when it did. It is not clear how the punishment - a permanent suspension - was decided on, nor how effectively it is likely to be enforced. It is not clear whether she or anyone else can contest the decision. Leaked emails from Facebook's handling of Alex Jones, a right-wing provocateur and conspiracy theorist accused of anti-Semitism, show similar confusion and arbitrary decision-making by nameless platform staff.¹¹ Matthew Prince, CEO of the security provider Cloudflare, reflected on his 2017 decision to remove safeguards for the neo-Nazi website Daily Stormer.

"I woke up this morning in a bad mood and decided to kick them off the Internet... It was a decision I could make because I'm the CEO of a major Internet infrastructure company... Literally, I woke up in a bad mood and decided someone shouldn't be allowed on the Internet. No one should have that power."^{12,13}

5. I. Loader, *Revisiting the Police Mission*, Police Foundation Insight Paper (April 2020), 8

6. Facebook Revenue and Usage Statistics (2020). Available at businessofapps.com/data/facebook-statistics/#5 (accessed October 2020)

7. G. De Gregorio, *Democratising Online Content Moderation: A Constitutional Framework*, Computer Law and Security Review (April 2020)

8. S. Zuboff, *Big Other: Surveillance Capitalism and the Prospects of an Information Civilization* (2015) 83

9. B. Schneier, *Data and Goliath* (2015) 58.

10. K. Langvardt, *Regulating Online Content Moderation*, *The Georgetown Law Journal* 106 (2018)

11. S. Kennedy, *Internal Facebook emails show struggle to crack down on anti-Semitism*, Channel 4 News. Available at <https://www.channel4.com/news/exclusive-internal-facebook-emails-show-struggle-to-crack-down-on-anti-semitism> (last accessed October 2020)

12. W. Oremus, *Cloudflare's CEO Is Right: We Can't Count on Him to Police the Internet*, *Slate* (August 2017). Available at slate.com/technology/2017/08/cloudflare-ceo-matthew-prince-is-right-we-cant-count-on-him-to-police-online-speech.html (last accessed October 2020)

13. It is also debatable whether "no one" should have that power.

The Daily Stormer, Hopkins and Jones all promote values that themselves run contrary to those of liberal democracy. They represent no great loss. Nevertheless, the process by which their speech has been moderated is consistent only with the unaccountable and arbitrary actions of a private company.

This arbitrariness and lack of clarity is mirrored in the experience of ordinary platform users. The most common complaint is a failure by platforms to take consistent action against content or behaviour that is supposedly breach of their terms of service. The Committee on Standards in Public Life's 2017 report on intimidation in public life highlighted a failure perceived by parliamentarians of platforms to act on the abuse they received.^{14,15} A ProPublica investigation from the same year found nearly half the content they reported to Facebook as hateful was mishandled, with damning anecdotal evidence for the inconsistencies in how the rules are applied.¹⁶ The Center for Countering Digital Hate's #WillToAct report highlighted a disparity between the claims platforms made about their action to tackle COVID-19 disinformation and its implementation.¹⁷ Similar inconsistency can be found in the haphazard moderation of historical photos of atrocities, documentation of police and state brutality, war reporting, satire and so on. Consistency and equality before law is a central tenet in both human rights legislation and in the policing mission. In practice, digital content moderation under a platform model is inconsistent and opaque.

The ability to challenge such a ruling is another keystone that is effectively nonexistent online, as is any ability to hold an individual or platform accountable for those rulings. Work by DotEveryone on *Better Redress* has highlighted shortfalls across the online world in effective redress for users, whether on social media, gaming, news websites or fraud, and have emphasised its particular scarcity for those users likely to need it the most.¹⁸ Just 28 percent of survey respondents to the *People Power Technology* survey felt they knew where to turn to when things went wrong online.¹⁹

The ability to challenge platform decisions is made more urgent by the sheer scale and complexity of the moderation process itself. The grey areas are enormous. Predictive classification of content is imperfect. A human content moderator must

make decisions in minutes, often about content in a language or from a context they do not understand. Mistakes are inevitable. Worse still (and similarly to the police officers with whom we draw parallels), platform moderators encounter a range of social problems and behaviours for which they may be poorly trained, unable to properly triage, or simply ill-suited to respond to. There are stories of content moderators tasked with responding to online threats of suicide and lacking the training, bandwidth or powers to do any more than simply scour the evidence from the platform. This sentence, taken from Bernardo Zacka's *When the State Meets the Street*, originally describes the police, but could equally well be written to describe content moderation teams.

*"They are condemned to being front-row witnesses to some of society's most pressing problems without being equipped with the resources or authority necessary to tackle these problems in any definitive way."*²⁰

Major platforms themselves have begun implementing routes for its users to challenge decisions, but these systems fail to remedy the problem: paths to redress are not independent, decisions are neither reviewable nor consistent, collective action is not supported, and decision-makers are unaccountable. The ability for citizens to challenge decisions made by authority is written into every level of the criminal justice system, from the supreme court down to independent police watchdogs and complaints commissions. Decisions taken can be guided by decades of recorded case law. The systems that make up the platform model are the polar opposite.

The importance of public accountability in law enforcement also comes through in literature around policing. While the police maintain a monopoly on legitimate force, their use of it must be subject to public scrutiny for it to remain legitimate.²¹ The top-down use of bans, blocks and content removal is a close approximation of force, and it is clear that its use under the platform model is not subject to any kind of public scrutiny.

SOLUTIONS IN PRINCIPLE

Online spaces require a radical overhaul of the principles they are built around, and consequently the principles they are regulated against. The debate around the compatibility of the platform

14. Intimidation in Public Life: A Review by the Committee on Standards in Public Life (2017)

15. Yvetter Cooper, Twitter. Available at <https://twitter.com/yvettercooperrmp/status/1121055948456583168?lang=en> (Last accessed October 2020)

16. A. Tobin et al., Facebook's Uneven Enforcement of Hate Speech Rules Allows Vile Posts to Stay Up, ProPublica, 2017. Available at: <https://www.propublica.org/article/facebook-enforcement-hate-speech-rules-mistakes> (Last accessed October 2020)

17. Center for Countering Digital Hate, #WillToActHow social media giants have failed to live up to their claims on the Coronavirus 'infodemic' (2020)

18. DotEveryone, Better redress: building accountability for the digital age, 2019. Available at <https://www.doteveryone.org.uk/wp-content/uploads/2019/12/Better-redress-evidence-review.pdf> (Last accessed October 2020)

19. Doteveryone People, Power and Technology: The 2018 Digital Understanding Report (2018)

20. B. Zacka, *When the State Meets the Street* (2017)

21. I. Loader, *Revisiting the Police Mission*, Police Foundation Insight Paper (April 2020), 4

model with liberal democracy will not end until decisions made around moderation by platforms are deemed legitimate. The argument that these are private companies taking private decisions has failed: these spaces are simply too socially and politically influential for states to let it go. To achieve legitimacy, moderation systems must be transparent, as must the individual and collective decisions made under those systems. The practice of platform moderation must be consistent, comprehensible and subject to challenge. From the perspective of regulation of online spaces, we believe that it is content moderation practices, not content, that should be the subject of government oversight.²²

SOLUTIONS IN PRACTICE

Improving content moderation in principle requires the injection of these tenets into the architectures of major platforms. Content moderation systems must be as transparent as possible and as consistently applied as possible. They must be subject to challenge, where possible through an independent channel, and their architects accountable for the decisions they make.

It is worth noting that where attempts to right wrongs have been successful, they have tended to come through two routes: litigation and investigative journalism. For instance, the Polish Panoptikon foundation filed a lawsuit in 2019 on behalf of a drugs education charity whose content had been removed. Dorota Głowacka, a lawyer at Panoptikon, explained the goal of the case:

*“is to challenge online platforms and incentivise them to move away from their current opaque and arbitrary methods of content moderation and to introduce measures which will better protect our freedom of speech... The user has to be informed why his or her content was blocked and be able to present arguments in his or her defence.”*²³

Governments looking to take steps to improve online spaces could do worse than support and protect investigative journalists, whistleblowers and civil society campaigns working to protect fundamental freedoms in the digital age.

One practical proposal we are sympathetic to is Giorgio De Gregorio’s proposed constitutional

framework, drawing a useful parallel between the use of constitutions to limit the powers of states, thereby giving those powers legitimacy.²⁴ Ironically perhaps, we believe that the state should be *promoting* its citizens rights to freedom of expression in the face of the autocratic, unaccountable private regimes that are sovereign in these spaces, rather than demanding further censorship for a list of online harms that gets longer every day. In essence, states should be regulating the ways platforms organise, moderate and police the content on those platforms to ensure those processes are aligned with democratic values.

De Gregorio proposes regulation in three parts. First, a notice system through which users are alerted when content they post or flag enters the moderation system, and can track their case through that process. Second, an explanation. Third, an opportunity to fight the case.

The proposal grounds the content moderation process in transparency. It must be clear to a user when their behaviour is policed. Under offline protocols, a person under arrest must be *told* they are under arrest. They must also be informed that it is the police arresting them, and what it is they are being arrested for. Following this example, it should be made clear to users of a platform how and why their content is being moderated, both with respect to their individual behaviour and content moderation practices in general. A user reporting content must have clarity over why their report was acted on or not. In his now famous 2006 lecture on *The Rule of Law*, Lord Bingham insisted “the law must be accessible and so far as possible intelligible, clear and predictable”.²⁵ Moderation guidelines for a platform like Facebook should be public knowledge, not the Kafkaesque subject of leaks and exposés.²⁶

Equally transparent should be the process itself. Algorithmic transparency and data processing transparency are at the heart of calls by civil society, governments and regulators. These calls tend to be rejected by platforms on grounds of economic interest, or, more dubiously, that algorithms are so mysterious and unknowable that it would be a waste of time.²⁷ Proposals for best-practice frameworks have been put forward by a range of national and international bodies, including a comprehensive review by the European Union, which supports increased regulatory oversight

22. K. Langvardt, Regulating Online Content Moderation, *The Georgetown Law Journal* 106 (2018)

23. S. Stolton, Facebook hit by landmark censorship lawsuit in Poland, *Euractiv*, 2019. Available at <https://euractiv.com/section/digital/news/facebook-hit-by-landmark-censorship-lawsuit-in-poland/> (Last accessed October 2020)

24. G. De Gregorio, Democratising Online Content Moderation: A Constitutional Framework, *Computer Law and Security Review* (April 2020)

25. Lord Bingham, *The Rule of Law*, 2006. Available at <https://www.cpl.law.cam.ac.uk/sir-david-williams-lectures/rt-hon-lord-bingham-cornhill-kg-rule-law> (Last accessed October 2020)

26. BBC, Leaks 'expose peculiar Facebook moderation policy', *BBC*, 2017. Available at <https://www.bbc.co.uk/news/technology-39997579> (Last accessed October 2020)

27. The Panoptikon Foundation has a good write up of this dubious position, found at: A. Forzyciaz, *Black-Boxed Politics: Opacity is a Choice in AI Systems*, Panoptikon Foundation. Available at en.panoptikon.org/articles/black-boxed-politics-opacity-choice-ai-systems (Last accessed October 2020)

backed by changes in procurement practices to begin to move the dial towards greater transparency.²⁸ In principle, at least, there appears to be growing consensus that opening up the 'black box' should be a regulatory aim.

Once a user has been notified of the content moderation process and been informed as to how and why a decision has been made, there must be an opportunity to challenge that decision. In principle, that challenge should be made through an independent institution, and systems of redress should allow for groups with collective complaints to bring them together. There are few examples of this being implemented online, though a number of Twitch channels now have collective sessions where users who have been banned are able to make their case and have it heard in front of a jury of their peers and the channel administrators for their reinstatement.

Part One underscores the gap between the principles and practice that underpin top-down moderation in the platform model, and those that underpin liberal democratic approaches to policing and order maintenance. Yet focusing on, or regulating, only the ways in which platforms handle their content moderation is barely half the equation. It ignores the power of the users, their communities, and those who lead those communities. It risks missing perhaps the most important factor in moderating and shaping positive online spaces.

28. A governance framework for algorithmic accountability and transparency, 2019. Available at [https://www.europarl.europa.eu/RegData/etudes/STUD/2019/624262/EPRS_STU\(2019\)624262_EN.pdf](https://www.europarl.europa.eu/RegData/etudes/STUD/2019/624262/EPRS_STU(2019)624262_EN.pdf) (Last accessed October 2020)

PART 2

COMMUNITY MODERATION

Part One looks to contrast our expectations of the rule of law and policing in a liberal democracy to our expectations of content moderation in the platform model. The analogy, although imperfect, helps build an understanding of why our mass transition to privately-owned online commons has felt at odds with how we understand the rules and enforcement of our behaviour. Its greatest strength, however, is in shining a light on the gaps. The platform model looks to preserve order through closely-controlled, unaccountable and top-down policing of content and behaviour. Offline, as we show below, this kind of policing is evidence of failure, not of success. Policing depends on social order. Social order is not the product of policing, yet this is the myth that underpins the vast majority of discussions of online regulation.

Social order, or a healthy community, depends on so much more than enforcement of the rules by force. Martin Innes positions modern-day conceptions of policing and social order in the historical context of the industrial revolution, but emphasises that it was one small part of the transformation of social and political institutions brought about by the change to the urban order. Institutions of public health, urban infrastructure programmes, social institutions, schools, unions, welfare systems and community and cooperative groups were all central to building up a new social order in the face of technological change. We believe that the digital platform revolution has not been met with similarly broad developments online. It is Ian Loader's belief that "there is no 'policing solution' to the problem of what makes societies secure and orderly."²⁹ If that is the case, why do we

default to content and behavioural policing as the window through which we look to make our online societies secure and orderly? Under the platform model, and certainly in discussions of platform regulation, scant attention is paid to these real drivers and how platforms might better build the architectures to support them.

What is the role of law enforcement in contributing to social order? To paraphrase David Leeney, modern conceptions of policing differentiate symptoms from cause. He identifies 'harms' as symptoms of a breakdown in this social order, and the role of the police as stepping in when the real drivers of social order and community health temporarily break down, or need a little time to work their magic. Martin Innes contrasts the "formal control" of law enforcement which plugs the gaps in the "informal control" that underpins a social order.

This is of particular importance given the range of responsibilities the police may have at one time or another. The prevalent community policing model demands far more of a police officer than kicking doors in. The 1962 Royal Commission on the Police notes the responsibilities of the police to include "befriending anyone who needs help" alongside law enforcement and the prevention of crime.³⁰ The Peelian principles emphasise the importance of willing cooperation between the police and their public and the need to prevent not just crime but disorder.³¹ Ian Loader is skeptical of the utility of any omnibus of police responsibilities, and fears the Peelian principles are little more than self-congratulatory, but for the purposes of this paper

29. I. Loader, *Revisiting the Police Mission*, Police Foundation Insight Paper (April 2020), 5
30. I. Loader, *Revisiting the Police Mission*, Police Foundation Insight Paper (April 2020), 2
31. *The Code of Ethics* (College of Policing), 5

they are useful reminders that policing does not stop at law enforcement. They are also reminders that the police often take on roles outside of any narrow stereotype. Mental health support is a prime example of one such task that brings with it exceptional challenges, intervention methods and training needs.

Order maintenance in our offline society depends on a web of interconnected institutions. For some, like the police, order maintenance is their primary reason for existing. For others, order maintenance is a byproduct of the work they do. At one end of the spectrum are the forces making up *plural policing*. Ian Loader explains:

*“What we might call a shift from police to policing has seen the sovereign state – hitherto considered focal to both provision and accountability in this field – reconfigured as but one node of a broader, more diverse network of power. Sure enough, this network continues to encompass the direct provision and supervision of policing by institutions of national and local government. But it now also extends ... to private policing forms secured through government; to transnational policing arrangements taking place above government; to markets in policing and security services unfolding beyond government; and to policing activities engaged in by citizens below government. We inhabit a world of plural, networked policing.”*³²

At the other end of the spectrum are the forces for whom order maintenance is a secondary or tertiary aim: mental health provision and other kinds of healthcare, social support, welfare, community leaders and organisations and so on. Together they make up the web of institutions that keep our societies healthy and orderly.

These lessons can be applied to the online world. First, that the platform model has turned to coercive models of policing as the primary method of order maintenance at the expense of devolving power and building a plurally policed space. Regrettably, this has likely been encouraged by government pressure to regulate on these lines. Second, that there are parts of the Internet where a powerful, positive social order has developed, and that the architecture underpinning that social order could and should be applied universally.

This architecture is defined by four principles: power, tools, incentives and dialogue. First, community policing depends on empowering actors and institutions outside of a centralised policing system. Second, those actors and institutions require the tools, routes and resources to make use of that power to shape and effect change in their communities. Third, those actors require some kind of incentive. This might be a belief that they are able to effect change, community recognition for their civic labour, or simply financial incentives. Finally, there is a need for genuine dialogue between the central authority, the wider group of actors and institutions responsible for social control, and the public. This dialogue is vital to evaluate the needs, priorities and expectations of the public and to ensure that their social values are mirrored in those charged with stewarding them.

THE PROBLEM IN PRACTICE

The hyper-centralisation of power in the platform model leaves no room for community policing. Twitter is probably the worst offender here, with YouTube and Facebook not far behind. An average user has no power in the platform model. An average platform user has no ability to shape the platform as a whole outside of reporting content and behaviour they disagree with to the central policing system and crossing their fingers. Work by scholars of policing and criminology helps us understand why this is problematic.

*“The contribution police make to security is deep in so far as police behaviour can and does provide individuals with a powerful token of their membership of a political community in ways that afford them the practical and symbolic resources required to manage, and feel relatively at ease with, the threats they encounter in their everyday life.”*³³

This dynamic is not supported by the models of moderation on major platforms.

Where tools designed to empower users exist in major platforms they err towards individual user responsibility: all major platforms now support users in blocking accounts or content they do not wish to see. But though this may be effective tools of individual empowerment, they do nothing for the health of the community as a whole. Where a

32. I. Loader, *Plural Policing and Democratic Governance*, *Social & Legal Studies* 9 3 (Sept 2020) 323–345

33. I. Loader, *Revisiting the Police Mission*, *Police Foundation Insight Paper* (April 2020), 15

community has seen a broader benefit, it is most often the result of coordination through non-platform tools. One example of this is the collation of Twitter block lists through which users build a communal pool of accounts to block.³⁴

More progress has been made in some aspects of platform design. A small number of users may become administrators or moderators under certain platform models such as Facebook Pages and Groups or subreddits on Reddit. Here, groups of users are significantly more empowered to impact the culture and behaviour of a community, and tools to support this are improving. Facebook's page administrators are now able to justify the removal of a post to the user who posted it, in line with the dialogue and transparency principles outlined above (though notably it does not allow a user to know who took the decision).

Similarly, Reddit moderators play a central role in determining the content and culture of the subreddits they administrate. One example of this can be seen below, again demonstrating the principles of dialogue and transparency.

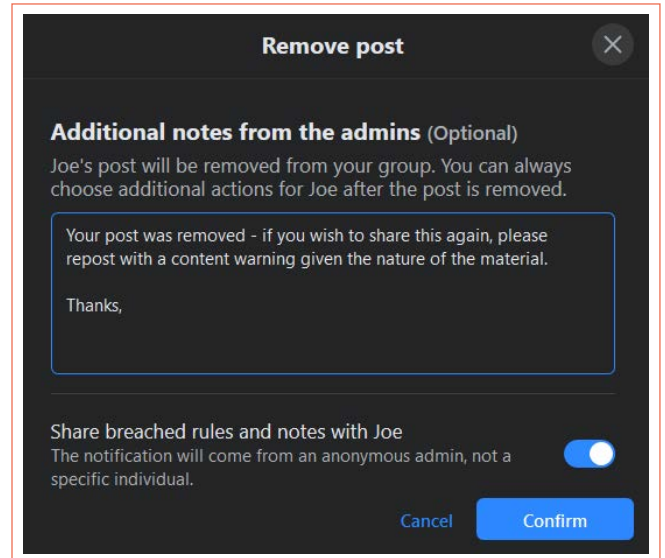


FIGURE 2.
FACEBOOK PAGE
CONTENT REMOVAL TOOL

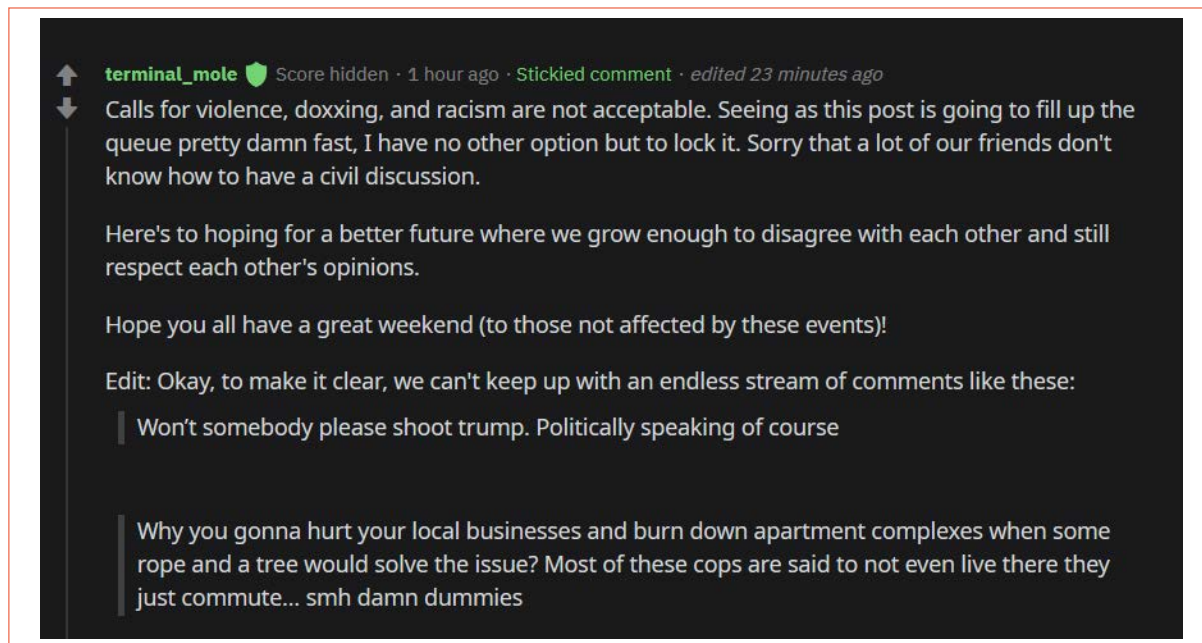


FIGURE 3.
A REDDIT MODERATOR EXPLAINS WHY
A DISCUSSION HAS BEEN SHUT DOWN

34. Block Together. Available at <https://www.theblockbot.com/> (Last accessed October 2020)

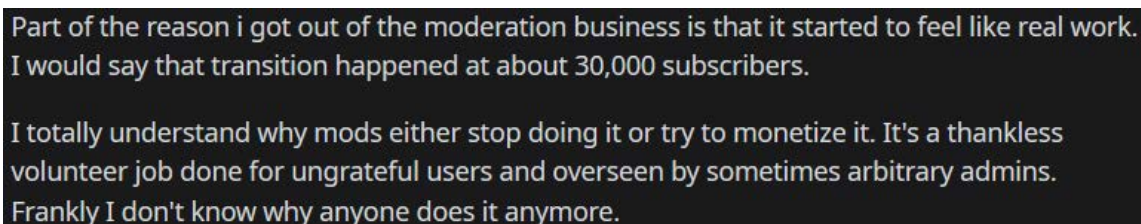
These systems are steps in the right direction but remain flawed. Administrators are largely unaccountable. Dedicated systems for electing, deselecting, reporting or challenging administrator decisions do not exist on Facebook pages. Administrators are frequently accused of abuses of power, including attempts to radicalise users, or of failure to provide sufficient stewardship. The power of 'super-moderators' on the platform is a subject of constant debate. One famous example saw a high-power Reddit moderator attempting to draw users into subreddits dedicated to conspiracy thinking and Holocaust denial.³⁵ This notwithstanding, the platform and its teams of moderators have built out extensive support and guidance for helping themselves function, including tools to report abuses of power and documentation and discussion to support good moderation practices.³⁶ By contrast, Facebook's guides are limited to explaining functionality.³⁷

Thirdly, there exist few formal incentives for community moderators. Interviewees disagreed over the importance of financial incentives in offline social control. Most suggested that the end goal

of a healthy society was the incentive in itself, and this observation likely carries over into the online world. There has been significant research into the motivations behind citizen moderators and how their labour should be understood. Nate Matias summarises the three positions: at any one time moderators are unwaged labour to the platform, online civic leaders to their community, or simply active participants in an elite club of web users.³⁸ Matias characterises the work as civic labour: negotiated volunteerism, but labour nonetheless. He notes the 2008 AOL settlement through which 14,000 unpaid moderators were awarded \$15 million after a class action lawsuit.

"No longer allowed the autonomy to imagine themselves as cultural gift-givers, the community leaders re-imagined themselves as mistreated employees and sued the company."

These volunteer moderators are the closest analogue we have to the prosocial institutions we depend on offline. Incentivising this civic labour and protecting moderators from harm should be a primary architectural goal of online platforms.



Part of the reason i got out of the moderation business is that it started to feel like real work. I would say that transition happened at about 30,000 subscribers.

I totally understand why mods either stop doing it or try to monetize it. It's a thankless volunteer job done for ungrateful users and overseen by sometimes arbitrary admins. Frankly I don't know why anyone does it anymore.

FIGURE 4.
A REDDIT MODERATOR VOICES THEIR
FRUSTRATIONS WITH VOLUNTEER
COMMUNITY ADMINISTRATION

35. Reddit. Available at https://www.reddit.com/r/self/comments/1xdwba/the_history_of_the_rkcd_kerfuffle (Last accessed October 2020)

36. For instance: [reddit.com/r/toolbox/](https://www.reddit.com/r/toolbox/), [reddit.com/r/ModSupport/](https://www.reddit.com/r/ModSupport/), [reddit.com/r/modclub/](https://www.reddit.com/r/modclub/)

37. Facebook, moderating your Facebook Page. Available at [facebook.com/facebookmedia/blog/moderating-your-facebook-page](https://www.facebook.com/facebookmedia/blog/moderating-your-facebook-page) or Facebook, Page Moderation Tips. Available at [facebook.com/business/a/page-moderation-tips](https://www.facebook.com/business/a/page-moderation-tips) (Last accessed October 2020)

38. J. Nathan Matias, The Civic Labor of Volunteer Moderators Online, *Social Media & Society*, (2019)

Finally, the platform model provides few systems supporting meaningful dialogue between moderators and the communities they police. This is most evident in interactions between users and platform-employed moderators and platform algorithms, where no dialogue is possible at all. David Leeney and Martin Innes both highlighted the importance of dialogue in any kind of neighbourhood or reassurance policing models, whereby those institutions and actors responsible for order maintenance work with communities to identify what matters to that community. Doing so helps set policing priorities, engages the community in prosocial thinking, and builds legitimacy in the policing process. Innes describes this as “community intelligence”:

*“By converting community engagement into a proactive and systematic task, and using the findings from this process to target interventions with a greater degree of precision than is commonplace in policing, a more concerted sense of direction and purpose is achievable.”*³⁹

This approach has merit when considering online spaces. It is quite clear that certain online spaces require different approaches to order maintenance than others. The Fossil Forum is never splashed across newspaper front pages for hosting extremist or violent content, or radicalising its user base.⁴⁰ Yet most platform engagement with communities to identify their priorities and expectations for online spaces is thin. The process most often takes place in the form of collective outrage channeled through the media and via PR firms, until a platform takes action.

SOLUTIONS IN PRINCIPLE

In summary, online spaces must undergo a series of changes to approach and architecture. Key to this is a devolution of power over a space to the communities who use them, and to those who perform the civic responsibilities of order maintenance. Those moderators who volunteer in this way should be incentivised, protected and rewarded. In return, systems must be built to ensure community members feel like active, politically- and socially-engaged participants in a space with power

to shape it and a responsibility for its contents. This turns on a continuing, structured dialogue between the platform, community leaders and volunteer moderators, and the community.

SOLUTIONS IN PRACTICE

There are examples of online practices and architectures that support community moderation significantly more effectively than under the current platform model.

Devolving power to the user base has been the foundation of the Wiki model, including Wikipedia. The barriers to participation on Wikipedia are extraordinarily low. A user does not even need to register to be able to make public changes to the site: by default, everyone has the power to shape the space.⁴¹ This model has been implemented outside of the main encyclopedia. Tens of thousands of Wikis have been set up to facilitate communal knowledge-sharing on subjects as diverse as eye health, sex work and cooking. Although Wikis are frequently the subject of vandalism and subterfuge, and the quality of pages does vary, studies support a view that major Wikis host broadly accurate information.⁴² Regardless, the model is a useful contrast to the default powerlessness experienced by users of most other platforms.

Other systems have experimented in user empowerment. Social networks built on blockchain systems such as Steemit and Minds reward users for participating on the platform through a token economy, and have spoken in the past of extending this into digital juries for content moderation.⁴³ *League of Legends*, an online game, has implemented workflows that gather user reports of negative behaviour, and should a critical mass be reached, users vote on whether to punish or pardon a reported user. An example of one such tribunal is shown below.

39. M. Innes et al., Neighbourhood Policing: The Rise and Fall of A Policing Model (2020) 198

40. The Fossil Forum. Available at Thefossilforum.com (Last accessed October 2020)

41. Wikipedia, Editing. Available at en.wikipedia.org/wiki/Help:Editing. (Last accessed October 2020)

42. For instance, R. Rosenzweig, Can History Be Open Source? Wikipedia and the Future of the Past. *The Journal of American History*, 93 (2006) 117–146. or L. H. Rector, Comparison of Wikipedia and other encyclopedias for accuracy, breadth, and depth in historical articles, *Reference Services Review*, 36 (2008) 7–22.

43. B. Ottman et al, Minds: The White Paper (2020). Available at <https://cdn-assets.minds.com/front/dist/en/assets/documents/Whitepaper-v0.5.pdf> (Last accessed October 2020)

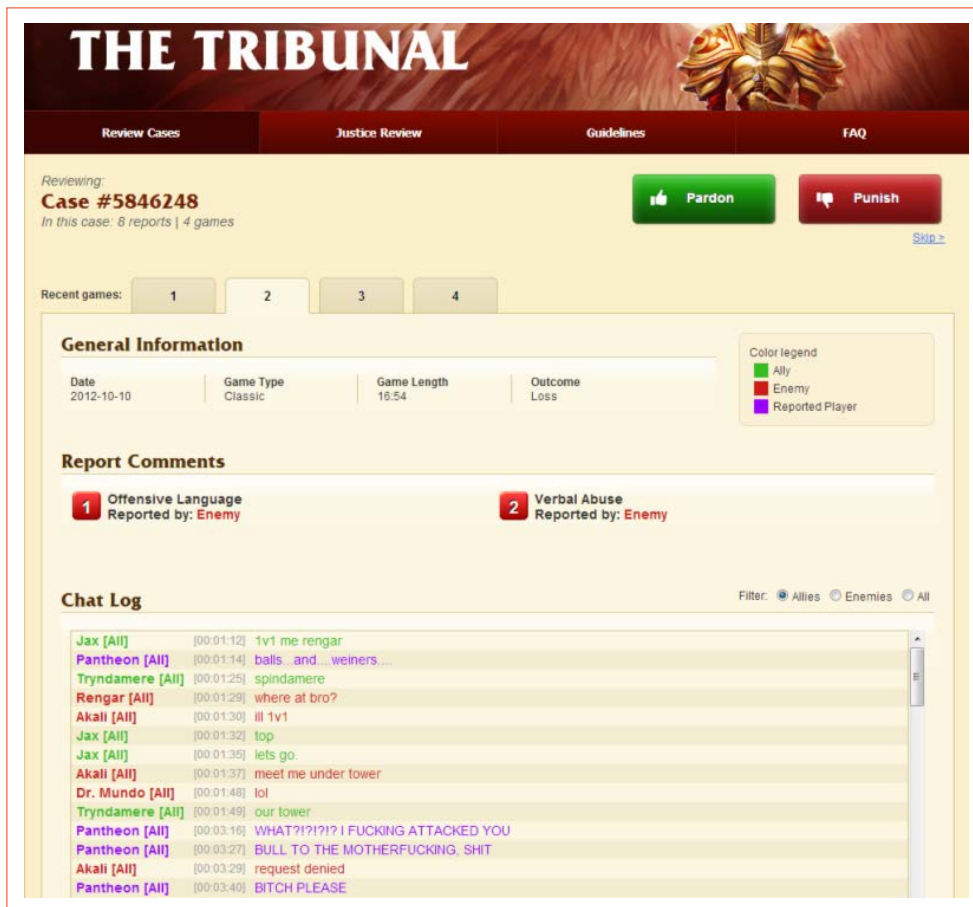


FIGURE 5.
A TRIBUNAL FOR A
PLAYER ENGAGING IN
ABUSE AND USING
OFFENSIVE LANGUAGE

Concluding their analysis of participants in this system, the researchers examining the system noted that:

“Explicit design support of communication and organization of judges and non-judges can facilitate the formation and development of social norms. For example, system design can make cases and judges public[ly] available to the community so that ... community members ... can view decision processes and recognize judges who contribute. Designers can also disclose their big data analytical results so that community members can quickly develop general understandings of community norms, instead of only getting to know what constitutes acceptable or unacceptable behavior only after receiving penalties.”⁴⁴

Research has suggested participants in digital juries found the systems “more procedurally just than existing common platform moderation practices”, and offline participation in jury systems results in an increased recognition of the legitimacy of the institutions of law enforcement and a sense of increased civic engagement.^{45,46} Alongside empowering users, distributed moderation is shown to be effective in moderating low-quality content and behaviour.⁴⁷

Users can also be empowered by helping choose or elect moderators. Stack Exchange, a network of 177 Q&A communities, holds regular elections to select community moderators, a process which includes virtual hustings.⁴⁸ This process has been replicated across a range of smaller forums, though formal support is limited and dependent on the decisions of existing authorities.⁴⁹

44. Y. Kou & X. Gui & S. Zhang & B. Nardi, Managing Disruptive Behavior through Non-Hierarchical Governance: Crowdsourcing in League of Legends and Weibo (2018)

45. J. Fan and A. X. Zhang, Digital Juries: A Civics-Oriented Approach to Platform Governance (2020) 1

46. V. P. Hans, J. Gastil, and T. Feller, Deliberative Democracy and the American Civil Jury (2014)

47. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Slash(dot) and burn: distributed moderation in a large online conversation space (2004)

48. Stackoverflow, 2017 Moderator Election. Available at <https://superuser.com/election/4> (Last accessed October 2020)

49. Diabetes.co.uk Moderator Elections 2019. Available at <https://www.diabetes.co.uk/forum/threads/moderator-elections-2019-voting-now-closed-thank-you-everyone-who-voted.164072?page-2> (Last accessed October 2020)

Finally, user empowerment can come through community-building tools. Reddit, Mastodon and Discord are all useful examples of platforms that allow the formation of micro-communities (though some number in the millions of participants) with their own sets of rules. Reddit, founded as “the front page of the Internet”, was first forced to split into ‘Safe for Work’ (SFW) and ‘Not Safe for Work’ (NSFW), and subsequently split into thousands of smaller communities, each with their own codes of conduct that sit below Reddit’s own. Mastodon follows a similar model. Users have power to shape their own spaces and hold participants in those spaces to more or less extensive expectations governing behaviour through running their own infrastructure.

Alongside an empowered user base, there are a range of solutions to incentivising moderators and reward them for carrying out their civic labour. Foremost here are reputational systems that reward visible benefits, flairs and badges to users who contribute to order maintenance. The vast majority of online forums support this, with users’ activity record, track record, years active and post count displayed next to their username. Examples of forum accolades for one user from Stack Overflow are shown below.

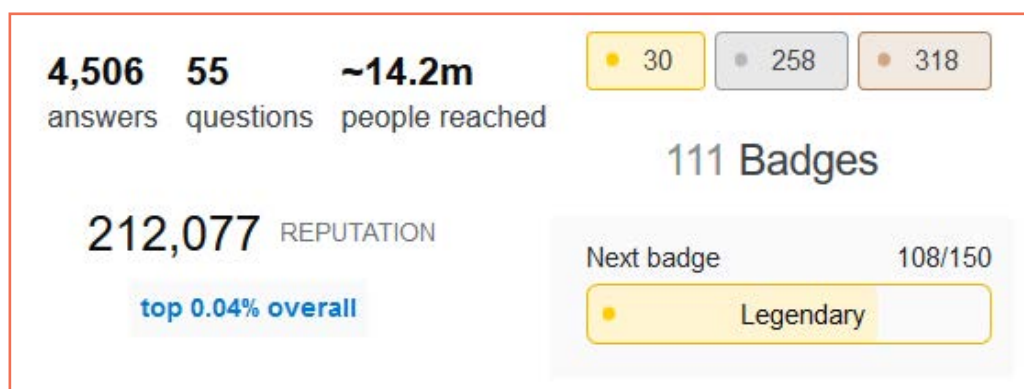


FIGURE 6.
USER LEVEL ACCOLADES FROM
STACKOVERFLOW, SHOWING BADGES,
REPUTATION AND IMPACT

Rewarding users for their volunteering reputationally has multiple benefits. For the user, it is recognition for their work, and in some cases reputational scores may be translated into other benefits, such as access to spaces, opportunities to vote or run for positions in the community, or other powers not available to the average user. It also positions the user as a good example to the wider community: achieving a reputation score of 212 thousand on StackOverflow is no mean feat. Where that user makes decisions on behalf of the community (for instance, by moderating content), their decisions are likely to be perceived as holding greater legitimacy. A further, positive side-effect of

reputation systems is the value it places on stable identities. This is explored in part 3 below, but in short, it incentivises users to place value in a given online which we see as a strong buffer against antisocial behaviour.

Reputation systems of this nature can be found across the digital world, from online games to anonymous chat rooms. In our view they represent powerful ways to incentivise pro-social behaviour.

Further incentives can be made by systematic improvements to the tools available to community administrators in carrying out their work. Across a large number of community-moderated

spaces, automatic moderation software has been implemented to tackle certain types of problematic online content *where appropriate*. A user spamming a subreddit would likely be picked up by an automated moderation tool, for instance, and Wikipedia is constantly crawled by dozens of bots that look to relieve contributors of the most menial tasks such as fixing broken links or updating dynamic lists.⁵⁰ In contrast to the platform model, this software is neither the default method nor the only method used to police the space. These tools, often built by the moderators who use them, using open, published code, instead allow community moderators to focus on those problems that most require human oversight, such as disputes between users or online harms without clear, machine-friendly attributes. One study has described this process as “liberating content moderators from the machinic role to which they are assigned, and treating them as protagonists of past, present, and future online cultures.”⁵¹

Finally, certain platforms have shown themselves to be capable of facilitating meaningful dialogue between users, community moderators and the rules and policing set by the platform itself. As noted above, authorities’ ability and willingness to listen and respond to the groups they police is an essential component of community policing. As with most changes to platform policies and systems, changes have tended to come in response to public outcries, rather than through measured and supported dialogue between platform and community. Grindr removed its ethnicity filter, StackOverflow changed its ways of working and Facebook removed advertising audiences for neo-Nazis after public and media pressure.^{52,53,54,55}

In summary, Part Two calls for platforms to empower users to be able to carry out civic labour online, provide them with the tools to do it, incentivise, reward, protect and listen to them.

50. See, for instance: <https://en.wikipedia.org/wiki/User:Cydebot>

51. See also: M. Ruckenstein, L. L. Maria Turunen, Re-humanizing the platform: Content moderators and the logic of care, *New Media & Society* (2019)

52. B. Hunte, Grindr removes ‘ethnicity filter’ after complaints, BBC (2020). Available at bbc.co.uk/news/technology-52886167 (Last accessed October 2020)

53. J. Hanlon, Stack Overflow Isn’t Very Welcoming. It’s Time for That to Change (2018). Available at stackoverflow.blog/2018/04/26/stack-overflow-isnt-very-welcoming-its-time-for-that-to-change/ (Last accessed October 2020)

54. K. Sutton, Facebook Let Advertisers Target Users Interested in Nazis, According to Report (2019). Available at adweek.com/digital/facebook-let-advertisers-target-users-interested-in-nazis-according-to-report/ (Last accessed October 2020)

55. BBC, Twitter apologises for letting ads target neo-Nazis and bigots, BBC, (2020) Available at bbc.co.uk/news/technology-51112238 (Last accessed October 2020)

PART 3

SELF-MODERATION AND ONLINE CULTURES AND NORMS

Below the rules, platform moderation teams and community moderators lies a final group of forces responsible for order maintenance: the social and cultural values and moral purposes of participants in a community. Social order does depend at least in part on a person's reluctance to risk sanction. But prosocial behaviour and compliance with the law depends as much, if not more, on some personal commitment to a community and to law-abiding behaviour.⁵⁶

The transition of our lives into online spaces appears to have tested how transferable these values are. We often hear how the anonymity, speed and low friction of online communication has turned us into trolls and monsters, willing to say and do things online that we wouldn't dream of doing face-to-face. This is not universally accurate. There are thousands of healthy, orderly, prosocial online communities. They count among their membership people who in the offline world might not be pegged as upstanding members of society or forces for social good, yet who online are engaged in building the public values of the spaces they inhabit. If further digitisation of our lives is an inevitability, which we think it is, we must identify how and why positive social values develop in some spaces and not others.

THE PROBLEM IN PRINCIPLE

The psychology, moral codes and patterns of behaviour that lead to antisocial behaviour online are daunting topics for a paper of this length. We explore the psychological effects which anonymity has on behaviour in an earlier report in this series.⁵⁷ To simplify the present discussion, we focus on one question: what can be learned from current and future platform architectures in encouraging prosocial behaviour? In answering this question, we approach three principles of platform architectures that are commonly associated with impacting on pro- or antisocial behaviour: anonymity and identity, friction, and measures of community identity.

To summarise, it has been argued that anonymity online, leading to disinhibition and a feeling of freedom from the consequences of a person's actions, increase the likelihood someone will act in an antisocial way.⁵⁸ It has been argued that a lack of friction in platform design increases the opportunity and reduces the barriers to antisocial behaviour. It has been argued that the wider the pool of values, expectations and norms among an online community, the more likely they are to clash and produce antisocial outcomes. There is likely some truth in all these positions, and they have useful lessons for informing the design of platforms going forward.

56. J. Jackson et al., *Why Do People Comply with the Law? Legitimacy and the Influence of Legal Institutions*, *British Journal of Criminology* 52 (2012) 2

57. J. Smith, E. Judson, E. Jones, *What's in a name?*, *Demos* (2020)

58. J. Suler, *The Online Disinhibition Effect*, *CyberPsychology & Behavior* 7, No. 3 (2004)

THE PROBLEM IN PRACTICE

Platform decisions on identity management and anonymity are frequently the subject of criticism. Clean Up the Internet in the UK cites dozens of MPs and civil society organisations critical of online anonymity. Conservative MP Phillip Davies is quoted: "It's common sense to most people that anonymity causes problems on social media." SNP MP John Nicholson suggests that "if platforms are serious about reducing the amount of vile abuse and dangerous misinformation on their platforms, they need to stop pretending anonymity isn't a problem."⁵⁹ Research is inconclusive. Some studies have pointed to anonymity being a factor when engaging in antisocial behaviour online, others have disputed this.⁶⁰

Less disputed is the role of friction. The argument is that it is so easy to act antisocially online that it is more likely to happen. By speeding up the technology by which we communicate and by reducing the barriers to a 'handshake' with another person or community, we weaken those persons' defenses against antisocial behaviour. There is certainly criminological support for this position: crime opportunity theory and the concept of opportunity crime foregrounds environmental and architectural design in enabling or preventing crime.⁶¹ The watchword of most major platforms has been 'connectivity', both as a system of economic growth and as a dubious moral imperative. Reflecting on a change in Facebook's mission statement Mark Zuckerberg described how

*"For the past 10 years, our mission has been to make the world more open and connected. We will always work to give people a voice and help us stay connected, but now we will do even more. Today, we're expanding our mission to set our course for the next 10 years. The idea for our new mission is: "bring the world closer together"."*⁶²

Which sounds like the same thing, frankly. Facilitating and speeding up the process by which people are able to interact with one another online is a core design principle for the platform model. In doing so, platforms have significantly reduced barriers to antisocial behavior. There are, however, success stories. Online support groups, Wikipedia's crowd-sourced encyclopedias, StackOverflow's community technology support all show that reducing friction does not in and of itself lead to

antisocial behavior. Opportunity crime still requires a criminal.

Connected to connectivity are measures of heterogeneity within online communities. Put enough people with differing value systems, behavioural norms and moral codes into a space and they will clash. Contrary to what the policy rhetoric of major platforms might tell you, connected people a community does not make. Community is about more than proximity or connection, but turns on shared values, purpose and a feeling of belonging.⁶³ Homogeneity of value, purpose and a shared sense of belonging underpin a successful community.

From the perspective of content moderation, this homogeneity impacts a community's ability to recognise an authority, set and respect rules, and work together on prosocial outcomes. Where McMillan and Chavis' elements of a "sense of community" do not exist, order maintenance is much more difficult.

A number of examples of platform architectural decisions already discussed in this paper can be seen in this light: many Reddit users did not want to see pornography on the platform, while many did, and the decision was taken to divide the community on those lines. Pressure to remove far-Right actors from major platforms is based on a rejection of one community's values by another, and the result has been a migration of far-Right groups to alternative platforms where those values are tolerated.⁶⁴

The chans (historically 4Chan and 8Chan) have historically been havens of behaviour and content that in most other parts of the Internet are banned: values like white supremacism, anti-Semitism and violent extremism, and behaviours like trolling, online intimidation and extremist radicalisation and recruitment. Moderation of these platforms reflects these community values. This is another paradox of the platform model. On the one hand, there is a commercial imperative to bring together and monopolise the provision of online communities. On the other is a rejection of the values that define many of those communities. Attempting to govern and moderate the moral codes, speech and behaviour of half the world's population under one Silicon Valley umbrella appears to be impossible.

59. Clean Up the Internet. Available at <https://www.cleantuptheinternet.org.uk/> (Last accessed October 2020)

60. For instance, J. Tremewan, Anonymity, Social Norms, and Online Harassment (2015) or <https://coralproject.net/blog/the-real-name-fallacy/>

61. M. Felson & R. Clarke, Opportunity Makes the Thief: Practical theory for crime prevention Police Research Series Paper (1998)

62. M. Zuckerberg, Bringing the World Closer Together, Facebook (2017). Available at [facebook.com/notes/mark-zuckerberg/bringing-the-world-closer-together/10154944663901634/](https://www.facebook.com/notes/mark-zuckerberg/bringing-the-world-closer-together/10154944663901634/) (Last accessed October 2020)

63. D. W. McMillan & D.M. Chavis, Sense of community: A definition and theory (1986) 16

64. Fieletz et al., Post-Digital Cultures of the Far Right: Online Actions and Offline Consequences in Europe and the US (2019) 12

SOLUTIONS IN PRINCIPLE

Part three focuses on individual agency. As such, solutions should begin with users having the opportunity and ability to hold and invest in an identity. Reputation, shame, and consequences of good or bad behaviour are meaningless without identity. Careful increases to friction should be considered. Finally, clarity and transparency over the rules, values, norms and moral codes of an online community should be central to platform design, and communities primarily conceived of in these terms.

SOLUTIONS IN PRACTICE

We believe that it is not anonymity that lies at the core of community disorder, but rather a failure of systems to promote and reward investment in a stable online identity. There are too many examples of healthy and prosocial online communities that allow users to participate without sharing their offline identities, and too many platforms on which people are antisocial under their real names, for anonymity to be the problem. What these spaces tend to have in common, however, is the ability for users to invest in a long-term, stable identity. Some examples of how this has been implemented are mentioned above: reputation systems, indicators of age or activity, or symbols of community-level recognition that reflect a user's contribution to a space. Without a stable identity, a user has nothing to lose in engaging in antisocial behaviour, but it does not then follow that their online identity must be tied to their offline one.

Take, for instance, the example of an avatar from a Massively Multiplayer Online Game (MMORPG). In these spaces, players invest many hours into a character, building up symbols of community value: whether that is a level indicator, a rank in a guild or group, or rare or expensive digital objects. Over time, this investment grows, and the risk of breaking the rules and losing it all increases. There is a significant body of academic literature that concludes there is a deep, nuanced and significant relationship between online avatars and their offline controllers.⁶⁵ By contrast, a freshly-made Twitter or Facebook account has no value whatsoever: breaking the rules and discarding that online identity comes at no cost to the person doing it. In our view, there is enough evidence supporting the implementation of support systems for stable identities on major platforms.

Rewarding moderators and community leaders with privileges and powers was discussed in Part Two, and the concept is extended here. Introducing friction, and reducing that friction as a reward, incentive or reflection of prosocial behavior is a common tactic that ought to be considered by major platforms. One simple example can be found on Reddit, where a new user does not necessarily have posting privileges, and must either wait or earn those privileges by participating in the community in some other way. 'Cooldown' times prevent spam, but also prevent a community being swamped or flooded by a handful of users. In online gaming, certain content can be made inaccessible until a player has carried out some kind of training, or simply spent enough time in the community to better understand its norms. One can easily imagine introducing a system on Twitter, for instance, that prevents a handshake between users before a new user has earned sufficient community capital. The dating app Bumble increases friction in a new direction: men cannot contact women until the woman has made the first move, a design decision the company explains improves the health of the platform for its users.⁶⁶ We believe that demand for social platforms that elegantly introduce friction will become challengers to existing frictionless models.

Finally, we believe that architectural decisions that prioritise McMillan and Chavis' four elements of community should be encouraged going forward. In practice, this can be as simple as clearly and transparently articulating the norms, rules and values that a given community identifies itself by. This has been standard practice on most online forums since usenet's message and bulletin boards, but on most major platforms this clarity has been lost as they battle with the scale and diversity of their enormous userbases. Finding ways to build coherent online communities with shared expectations has also led to innovative platform decisions. Breaking the rules in some Minecraft servers will result in a user being moved to an alternative server populated by others who broke the rules, both punishment and a solution for users who want their online spaces to operate under different rules. Anecdotally, we believe that where communities share a purpose - a place to discuss fishing or fossil hunting - the communities tend to be healthier than spaces where connections are made without any clear reason. This being said, more research is required on how to best measure the strength and health of a community, some of which will be carried out as part of the ongoing Good Web Project at CASM.

65. F. Sibilla & T. Mancini, I am (not) my avatar: A review of the user-avatar relationships in Massively Multiplayer Online Worlds, *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, 12 2 (2018)

66. wired.co.uk/article/bumble-whitney-wolfe-sexism-tinder-app (Last accessed October 2020)

PART 4

RECOMMENDATIONS

This short paper has been focused on presenting solutions in principle and practice to the challenges presented by content moderation practices and systems. We present the above recommendations in summary:

REGULATION & POLICY

1. With regards to legal but harmful content, regulation should target the systems and effectiveness of content moderation, rather than individual categories of content.
2. However, serious crime - exploitation of children, human trafficking and slavery - constitute problems where forceful intervention is justified and legitimate. Where technological solutions to fight these crimes can be shown to be effective, they should be implemented and enforced.

PLATFORM DESIGN

3. Platforms should move away from 'bolting on' fixes to fundamentally adjusting their systems of self-governance and user empowerment. This report touches on many examples, including increasing levels of friction, verifying aspects of their users identity (for instance, the Verification of Children Online work led by the UK government), provision and reward for stable identities, or improving a user's ability to impact the communities they live in.
4. Larger platforms should support internal and external research and development to understand how alternative models to top-down platform moderation may be effective in fighting online harms.

5. Platforms should recognise the civic labour performed by their users in policing or moderating their spaces, and incentivise and reward them for doing so.
6. Larger platforms should support international public messaging aimed at raising awareness of the opportunities and routes average Internet users may have in reshaping and improving their online communities, with the caveat that this depends on users having access to tools and systems that are genuinely empowering (see 3 above).

PLATFORM TERMS OF SERVICE

7. Platform terms of service must be made clear and comprehensible to the average user.
8. Platforms should work to improve transparency over the rules and processes that govern the spaces they own. This includes:
 - Clarity over permissible content and behaviour
 - Clarity over moderation practices
 - Clarity over a user's individual experience of a moderation decision
9. Platforms should work to ensure their terms of service are consistently applied.
10. Platforms should provide effective routes to redress for users who believe their behaviour has been incorrectly or unjustly policed.

RECOMMENDED READING

Given the scale and complexity of this subject, and the variety of perspectives that we believe are fruitful avenues for discussion, we recommend three short, accessible texts we found particularly helpful in guiding our own thinking.

Ian Loader, *Revisiting the Police Mission*
(Police Foundation Insight Paper April 2020)

Found at: https://policingreview.org.uk/wp-content/uploads/insight_paper_2.pdf

Giovanni De Gregorio, *Democratising online content moderation: A constitutional framework*

Found at: <https://www.sciencedirect.com/science/article/pii/S0267364919303851>

Kyle Langvardt, *Regulating Online Content Moderation*

Found at: <https://www.law.georgetown.edu/georgetown-law-journal/wp-content/uploads/sites/26/2018/07/Regulating-Online-Content-Moderation.pdf>

DEMOS

PUBLISHED BY DEMOS OCTOBER 2020
© DEMOS. SOME RIGHTS RESERVED.
15 WHITEHALL, LONDON, SW1A 2DD
T: 020 3878 3955
HELLO@DEMOS.CO.UK
WWW.DEMOS.CO.UK