IMPOSING LIABILITY ON ONLINE INTERMEDIARIES FOR VIOLENT USER-GENERATED CONTENT: AN AUSTRALIAN PERSPECTIVE

Emily Irwin[*] & Niloufer Selvadurai, Ph.D.[**]

Cite as: Emily Irwin & Niloufer Selvadurai, *Imposing Liability on Online Intermediaries for Violent User-Generated Content: An Australian Perspective*, 28 RICH. J.L. & TECH. 1 (2021).

[*] Emily Irwin, BSocSc LLB (Hons I) Macq, Attorney-General's Department, Legal Policy, Canberra, Australia. This article was written prior to the author's engagement with the Attorney-General's Department. The views expressed in this paper are wholly the personal views of the author and do not reflect the views of the Attorney-General's Department or the Australian Government.

[**] Professor Niloufer Selvadurai, BA LLB (Hons I) USyd, PhD Macq, Macquarie Law School, Macquarie University, Sydney, Australia.

ABSTRACT

The Australian 2019 *Criminal Code Amendment (Sharing of Abhorrent Violent Material) Act* requires online intermediaries, including social media services and digital platform providers, to take steps to remove abhorrent violent material that is accessible in Australia. While the Act represents a significant landmark in internet governance, it is questionable whether it appropriately aligns the need to remove violent material with the need to support legitimate socially-beneficial online speech. In this context, this paper critically analyzes the operation of the Act, considering whether and to what extent compliance with the Act may lead to the over-removal of socially-beneficial and lawful speech on social media. Beyond Australian law, this paper considers the broader discourse of international law in this important and evolving area of law.

I. Introduction

[1] As the volume of violent material shared on social media continues to increase, a critical legal issue to be addressed is the extent to which online intermediaries should be liable for such user-generated content. In 2019, Australia passed the *Criminal Code Amendment (Sharing of Abhorrent Violent Material) Act 2019* (Cth) ("AVM Act"). It requires "content services" (including social media services) and "hosting services" (including digital platform providers) to "expeditiously" remove "abhorrent violent material" ("AVM") capable of being accessed in Australia.[1] "AVM" is defined as audio or video recorded by a perpetrator or accomplice, depicting terrorism, rape, torture, murder, attempted murder or kidnapping.[2] Significantly, the fault element for this offence is recklessness, and the maximum penalties are extremely severe so as to encourage compliance through deterrence.[3] This approach was designed to ensure intermediaries engage in active content removal, but limited only to "the worst types of material that can be shared online".[4] However, despite its potentially significant effect, the AVM Act has been the subject of limited scholarly analysis. In such a context, the objective of this paper is to critically analyze the AVM Act and consider whether it should be amended to more appropriately regulate the sharing of AVM on social media. In considering what constitutes appropriate regulation in this area, the paper will apply the utilitarian principle of criminalization ("UPC"), which postulates that an act should only be criminalized in a particular manner, such as through the AVM Act, if it would maximize overall utility and produce a net social

[1] *Criminal Code Amendment (Sharing of Abhorrent Violent Material) Act 2019* (Cth) ss 474.34(1)–(8) (Austl.) [hereinafter *Abhorrent Materials Act*].

[2] *Abhorrent Materials Act*, *supra* note 1, at ss 474.31, 474.32(1).

[3] *Id.* at ss 474.34(4), (8). *But see id.* at s 474.37 (noting the AVM Act offers defenses to preserve public interest uses of AVM).

[4] Explanatory Memorandum, Criminal Code Amendment (Sharing of Abhorrent Violent Material) Bill 2019 (Cth) paras 49, 56 (Austl.); Commonwealth, *Parliamentary Debates*, House of Representatives, 4 April 2019, 1849–50 (Christian Porter, Attorney-General) (Austl.).

benefit.[5] The paper will use the UPC framework to provide criteria to critique the AVM Act, weigh competing interests and develop recommendations for law reforms to maximize social benefit.

[2] Australia's AVM Act aligns with broader cybercrime and online content regulation trends by criminalizing the actions of online intermediaries, rather than individual users. Unlawful online content is increasingly policed through "public-private cooperation," in which private internet intermediaries are legally required to monitor online content on behalf of their governments.[6] Additionally, social media services increasingly conduct "privatized censorship" by removing speech on their platforms according to their terms of service, with minimal public oversight.[7] The AVM Act entrenches privatized censorship by mandating that platforms *should* remove AVM, but not specifying *how* they should do so.[8] Thus, the AVM Act effectively imposes both state-sanctioned

[5] Thomas Søbirk Petersen, *A Soft Defense of the Utilitarian Principle of Criminalization*, 26 RES PUBLICA 123, 126 (2020); Geraldine Szott Moohr, *Defining Overcriminalization Through Cost-Benefit Analysis: The Example of Criminal Copyright Laws*, 54 AM. U. L. REV. 783, 786 (2005). *See generally* JEREMY BENTHAM, AN INTRODUCTION TO THE PRINCIPLES OF MORALS AND LEGISLATION 134–35 (Batoche Books 2000) (1781) (asserting that a law fails to uphold the UPC if its negative consequences outweigh the social benefits it produces).

[6] Hannah Bloch-Wehba, *Automation in Moderation*, 53 CORNELL INT'L L.J. 41, 46 (2020); Jack M. Balkin, *Free Speech is a Triangle,* 118 COLUM. L. REV. 2011, 2017–18 (2018) [hereinafter *Free Speech is a Triangle*]; Niva Elkin-Koren & Maayan Perel, *Guarding the Guardians: Content Moderation by Online Intermediaries and the Rule of Law*, *in* OXFORD HANDBOOK OF ONLINE INTERMEDIARY LIABILITY, 669, 673 (Giancarlo Frosio, ed., 2020) [hereinafter *Guarding the Guardians*].

[7] Jack M. Balkin, *Old-School/New-School Speech Regulation*, 127 HARV. L. REV. 2296, 2298–99, 2309 (2014) [hereinafter *Old-School/New-School*]; Bloch-Wehba, *supra* note 6, at 46–47; Seth F. Kreimer, *Censorship by Proxy: The First Amendment, Internet Intermediaries, and the Problem of the Weakest Link,* 155 U. PA. L. REV. 11, 27–28 (2006).

[8] Bloch-Wehba, *supra* note 6, at 45.

privatized censorship and public-private collaboration to achieve enforcement and reduce AVM on social media.

[3] However, while a variety of scholars have argued that public-private cooperation, enforced by strict criminal punishment, can potentially reduce unlawful online content, this analysis has not been extended to the merits of the AVM Act. Internet intermediaries are easy to identify and are best positioned to directly block unlawful content.[9] Hence, a variety of authors have drawn on Deterrence Theory to argue that strict criminal punishment heavily induces compliance by rational corporations,[10] yet none have applied this reasoning to examine the AVM Act's ability to enforce AVM removal.

[4] Moreover, while there is scholarly discourse on the deleterious effects of the over-removal of legitimate material on social media [11] and the moderation of harmful content, these issues have not been examined in detail within the context of the AVM Act.[12] Without a codified Australian right to free speech[13], authors analyzing the AVM Act must find other reasons to explain *why* over-removal of legitimate speech is concerning.

[9] *Free Speech is a Triangle*, *supra* note 6, at 2019–20; Niva Elkin-Koren & Eldar Haber, *Governance by Proxy: Cyber Challenges to Civil Liberties*, 82 BROOK. L. REV. 105, 113–14 (2016) [hereinafter *Governance by Proxy*]; Aniket Kesari et al., *Deterring Cybercrime: Focus on Intermediaries*, 32 BERKELEY TECH. L.J. 1093, 1098 (2017).

[10] John T. Byam, *The Economic Inefficiency of Corporate Criminal Liability*, 73 J. CRIM. L. & CRIMINOLOGY 582, 586 (1982); Raymond Paternoster & Sally Simpson, *Sanction Threats and Appeals to Morality: Testing a Rational Choice Model of Corporate Crime*, 30 L. & SOC'Y REV. 549, 553, 579–80 (1996).

[11] *Free Speech is a Triangle*, *supra* note 6, at 2030–31.

[12] Evelyn Douek, *Australia's "Abhorrent Violent Material" Law: Shouting "Nerd Harder" and Drowning Out Speech*, 94 A.L.J. 41, 47–48 (2020); Bloch-Wehba, *supra* note 6, at 45, 70, 88; Kreimer, *supra* note 7, at 27–28.

[13] *Freedom of information, opinion and expression*, AUSTRALIAN HUM. RTS. COMM'N, https://humanrights.gov.au/our-work/rights-and-freedoms/freedom-information-opinion-and-expression [https://perma.cc/XSP9-P4GP].

Existing literature on the AVM Act considers the issue of over-removal, but it does not explicitly consider whether this is an acceptable trade-off to achieve the law enforcement objectives of the AVM Act.[14] Little of the discourse that considers potential reform measures to ensure effective and accountable moderation of harmful content online extends to analyzing the merits of the moderation system enacted by the AVM Act. Common recommendations include transparency reporting and mandatory appeals processes.[15] For instance, when examining the AVM Act, Douek recommends non-criminal regulations that incentivize social media platforms to implement appropriate moderation systems, rather than punishing failure to remove content,[16] but stops before explicitly identifying what an appropriate moderation system should entail, or who would implement those requirements.[17] Finally, there is some discord as to which solution is most appropriate.[18]

[5] This paper seeks to advance discourse in this area by providing a close critical analysis of the operation of AVM Act. Section II begins by analyzing the AVM Act's enforcement mechanisms and considers the extent to which the AVM Act's objective of inducing compliance through public-private cooperation and severe criminal sanctions incentivizes the removal of AVM. In doing so, this paper considers the extent to which social media and hosting services are amenable to deterrence, and considers whether the AVM Act's uncertain enforcement may limit deterrence by identifying specific categories of content and hosting services who are less likely to comply. Extending this analysis, Section III examines whether and to what extent compliance with the AVM Act may lead to the incidental

[14] Douek, *supra* note 12, at 51.

[15] *E.g.*, Danielle Keats-Citron, *Extremist Speech, Compelled Conformity, and Censorship Creep*, 93 NOTRE DAME L. REV. 1035, 1067–68 (2018); Bloch-Wehba, *supra* note 6, at 83, 87–88, 91; Douek, *supra* note 12, at 52, 59.

[16] Douek, *supra* note 12, at 52.

[17] *Id.*

[18] *See, e.g.*, *Guarding the Guardians*, *supra* note 6, at 674 (questioning the usefulness of mandatory transparency reporting).

over-removal of socially-beneficial speech on social media. In doing so, this paper considers whether social media platforms' current content moderation capabilities are congruent with the moderation standard required to adhere to the AVM Act, and considers whether such potential incongruencies could lead to collateral over-removal of lawful speech. Beyond domestic legal considerations, Section IV situates the AVM Act within broader international law reform discourse. Finally, based on these deliberations, Section V of this paper presents a variety of recommendations for reform to ensure that the AVM Act maximizes social benefit by achieving its stated objectives while reducing negative externalities.

II. The Merits of the AVM Act's Enforcement Mechanisms

[6] The AVM Act contains two distinct legal measures for enforcing the removal of AVM. First, the Act imposes intermediary liability by targeting content and hosting services who fail to remove AVM, rather than the individuals uploading the content.[19] Ideally, this will make the AVM Act more practical to enforce. Second, it seeks to achieve deterrence by imposing harsh criminal penalties on those who fail to comply.[20] Theoretically, these elements should combine to strongly incentivize content and hosting services to actively monitor and "expeditiously" remove AVM. In doing so, the AVM Act should produce significant social utility by reducing the incidence of harmful AVM online. However, as this paper will discuss, the expected enforcement benefits may be lower than legislators anticipated.

A. Imposing Liability on Online Intermediaries

[7] Recognizing that the internet is an open network in which almost any individual user has the power to share content with a worldwide audience within seconds,[21] the AVM Act uses public-private collaboration

[19] Douek, *supra* note 12, at 43.

[20] *Abhorrent Materials Act*, *supra* note 1, at ss 474.33(1), 474.34(9)–(10)(a).

[21] *Governance by Proxy*, *supra* note 9, at 110.

as an enforcement tool by criminalizing intermediaries who fail to expeditiously remove AVM. This carries significant benefits, including easier detection and removal of AVM, and theoretically, easier prosecution in the case of a transgression.[22] Content and hosting services create the infrastructure necessary to facilitate the sharing of AVM on social media, making them "ideal partners" for enforcement.[23] The AVM Act targets both content and hosting services, effectively providing two layers at which AVM can be detected and removed.[24] Content services, such as Facebook, monitor users' activity to block and remove content.[25] Most already undertake extensive moderation to remove content that breaches their terms of service.[26] Furthermore, hosting services have the ability to cease hosting entire websites, evidently disrupting their ability to remain online.[27] Accordingly, even if a content service website fails to remove AVM, their hosting service could withhold the infrastructure necessary for the content service to remain operational, thwarting the spread of the AVM altogether.[28] By shifting enforcement efforts to target these intermediaries instead of the

[22] *Abhorrent Materials Act*, *supra* note 1, at s 474.34.

[23] *Governance by Proxy*, *supra* note 9, at 113.

[24] *Abhorrent Materials Act*, *supra* note 1, at ss 474.39, 474.30, 474.33.

[25] *Governance by Proxy*, *supra* note 9, at 113; Terry Flew et al., *Internet regulation as media policy: Rethinking the question of digital communication platform governance*, 10 J. DIGIT. MEDIA & POL'Y 33, 45 (2019); ROBERT G. PICARD & VICTOR PICKARD, UNIV. OXFORD, ESSENTIAL PRINCIPLES FOR CONTEMPORARY MEDIA AND COMMUNICATIONS POLICYMAKING 6 (2017).

[26] Tarleton Gillespie, *Regulation of and by Platforms*, *in* THE SAGE HANDBOOK OF SOCIAL MEDIA 266 (Jean Burgess, Alice Marwick & Thomas Poell eds., 2018); Majid Yar, *A Failure to Regulate? The Demands and Dilemmas of Tackling Illegal Content and Behavior on Social Media*, 1 INT'L J. CYBERSECURITY INTEL. & CYBERCRIME 5, 12 (2018); Flew et al., *supra* note 25, at 48.

[27] Douek, *supra* note 12, at 43, 51.

[28] *See id.* at 51 ("8chan is having difficulty remaining online after its hosts pulled their services when the forum was the site of the advance announcement of three mass shootings in six months.")

individuals who upload AVM, the AVM Act effectively utilizes their unique content removal capabilities.[29]

[8] The shift in enforcement is crucial, because law enforcement simply does not have the same level of access, resources, or technical capacity to perform the monitoring and content removal necessary to control the spread of AVM.[30] On YouTube alone, users upload 400 hours of video every minute.[31] It is unreasonable to expect law enforcement to monitor this volume of content for every social media platform, given their limited technical capacity and resources.[32] By criminalizing intermediaries, the AVM Act avoids this enforcement issue.

[9] Further, prosecuting each individual who uploads AVM is not practically feasible given the scale of content uploaded online.[33] The Christchurch attack video elucidates a pertinent example.[34] In that case, the original perpetrator's livestream was re-uploaded to Facebook over 1.5 million times by different users.[35] Given this large scale, law enforcement would not be able to viably detect each of these videos and identify each

[29] *Free Speech is a Triangle*, *supra* note 6, at 2019–20; Doug Lichtman & Eric C. Posner, *Holding Internet Service Providers Accountable*, 14 SUP. CT. ECON. REV. 221, 235–37 (2006).

[30] *Free Speech is a Triangle*, *supra* note 6, at 2019–20; Jason H. Peterson et al., *Global Cyber Intermediary Liability: A Legal & Cultural Strategy*, 34 PACE L. REV. 586, 598 (2014); *Governance by* Proxy, *supra* note 9, at 113–14.

[31] Flew et al., *supra* note 25, at 41.

[32] *See* Yar, *supra* note 26, at 12; Jack M. Balkin, *Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation*, 51 U.C. DAVIS L. REV. 1149, 1175 (2018) [hereinafter *Free Speech in the Algorithmic Society*].

[33] *Free Speech is a Triangle*, *supra* note 6, at 2020.

[34] Douek, *supra* note 12, at 41–42.

[35] *Id.* at 41.

individual offender without cooperation from intermediaries.[36] Moreover, even if each individual was identifiable, the government would certainly encounter insurmountable resource constraints when prosecuting each offender, particularly if they are located in different jurisdictions.[37] In turn, individual offenders would likely perceive this low certainty of punishment in a way that would lessen the deterrent effect.[38] Conversely, the AVM Act's choice to target intermediaries narrows the scope to easily-identifiable entities, ideally making enforcement a much simpler task and theoretically increasing the certainty of punishment.[39]

[10] However, despite these benefits, the AVM Act specifically criminalizes the failure to remove AVM, placing the challenging enforcement task back onto the government.[40] To make the determination that an offense has been committed, law enforcement must determine whether or not social media and hosting services are actually removing AVM. This inevitably involves some degree of monitoring the content uploaded to social media to determine compliance,[41] a highly impractical

[36] *See generally Free Speech is a Triangle*, *supra* note 6, at 2019–20 (discussing reasoning behind private-public cooperation in regulating internet speech); Yar, *supra* note 26, at 11–12 (stating that monitoring and regulation by social media platforms in conjunction with efforts by public law enforcement agencies is necessary to bridge the capacity gap).

[37] *See Free Speech is a Triangle*, *supra* note 6, at 2020.

[38] S*ee generally* Mark C. Stafford, *Deterrence Theory: Crime*, *in* 6 INTERNATIONAL ENCYCLOPEDIA OF THE SOCIAL & BEHAVIORAL SCIENCES 255, 255–56 (James D. Wright ed., 2d ed. 2015); Raymond Paternoster, *How Much Do We Really Know About Criminal Deterrence?*, 100 J. CRIM. L. & CRIMINOLOGY 765, 767–72, 781–82, 784, 787 (2010); CESARE BECCARIA, ON CRIMES AND PUNISHMENTS 42–43, 93–94 (Henry Paolucci trans., 1963); BENTHAM, *supra* note 5, at 19–20, 117, 137.

[39] *Free Speech is a Triangle*, *supra* note 6, at 2020; *Old-School/New-School*, *supra* note 7, at 2338.

[40] Douek, *supra* note 12, at 42–43.

[41] *Free Speech is a Triangle*, *supra* note 6, at 2020; *Old-School/New-School*, *supra* note 7, at 2304–05.

and challenging task suggesting that transgressions will likely go unnoticed.[42]

[11] This problem could potentially be remedied by mandating that social media and hosting services report on their content removal practices, or introducing some other form of transparency requirement similar to those in other jurisdictions.[43] However, the AVM Act does not require social media platforms to report on their content removal practices.[44] While intermediaries must report abhorrent violent conduct occurring within Australia, this does not provide any assistance if an incident occurs overseas, and provides no insight into intermediaries' actual removal practices.[45] Without a legal requirement to report on their content removal practices, the AVM Act may incentivize platforms to cover up any potential system failures to avoid prosecution, exacerbating these enforcement difficulties.[46]

[12] One counterargument is that a failure to remove AVM will generally be obvious to law enforcement, so the incentivization problem is an exaggerated concern. This point has some merit. For example, in widely publicized cases such as the Christchurch attack video, failure to expeditiously remove AVM will be extremely obvious.[47] Yet given the vast

[42] *See, e.g.*, Yar, *supra* note 26, at 5–6; *Free Speech in the Algorithmic Society*, *supra* note 32, at 1175.

[43] *See, e.g.*, Netzdurchsetzunggesetz [NetzDG] [Network Enforcement Act], Oct. 1, 2017, BUNDESGESETZBLATT [BGBl] at I 3352 (Ger.) (regulating reporting of and handling of complaints about unlawful content); DEPARTMENT FOR DIGITAL, CULTURE, MEDIA AND SPORT, ONLINE HARMS WHITE PAPER: FULL GOVERNMENT RESPONSE TO THE CONSULTATION, 2020, Cm. 354, at 3–4 (UK) [hereinafter FULL GOVERNMENT RESPONSE] (broadly describing the UK's regulatory framework).

[44] Douek, *supra* note 12, at 44–45.

[45] *Abhorrent Materials Act*, *supra* note 1, at s 474.33.

[46] Bloch-Wehba, *supra* note 6, at 53.

[47] Douek, *supra* note 12, at 41–42.

amounts of content being uploaded online, other atrocities are inevitably ignored. In July 2019, after the AVM Act came into force, pictures of 19 year old Bianca Devins' horrific murder were uploaded to Instagram.[48] The images were not removed "expeditiously" and remained online for several days.[49] However, when the eSafety Commissioner was questioned on whether they would issue Instagram a removal notice, they stated they had not received any complaints and appeared unaware of the incident.[50] As Douek notes, the only difference between the pictures of Devins' murder and the Christchurch attack video was the amount of public attention and media coverage they received in Australia.[51] This oversight exemplifies the infeasibility of expecting law enforcement to actively and stringently monitor whether content is being removed "expeditiously" without adequate transparency or reporting requirements.[52] Without implementing any transparency requirement, the current approach makes law enforcement's job unnecessarily difficult. Inevitably, only the most publicized cases will result in prosecution, which is not conducive to maximum enforcement.[53]

B. Criminalization to Incentivize the Removal of AVM by Online Intermediaries

[13] While enlisting intermediaries to remove AVM is necessary for enforcement, it will be fruitless if the AVM Act does not adequately

[48] *Id.* at 56.

[49] *Id.*

[50] *Id.*

[51] *Id*. at 57.

[52] Mark MacCarthy, *Transparency Requirements for Digital Social Media Platforms: Recommendations for Policy Makers and Industry*, *in* TRANSATLANTIC WORKING GRP. ON CONTENT MODERATION ONLINE AND FREEDOM OF EXPRESSION 5, 13 (2020), https://www.ivir.nl/publicaties/download/Transparency_MacCarthy_Feb_2020.pdf. [https://perma.cc/LW7W-3R22].

[53] *See* Douek, *supra* note 12, at 41–42.

incentivize platforms to actually carry out this removal task in practice. The AVM Act is a criminal law, seeking to incentivize compliance through deterrence and harsh criminal punishment, particularly by imposing large fines.[54] Although not explicit, this assumes online intermediaries are rational profit-maximizers who will choose to comply with the AVM Act if criminal sanctions are sufficiently harsh so as to outweigh the benefits of offending.[55] Most major social media and hosting services are corporations, meaning they are, at least to some degree, driven by rational profit-maximization.[56] While some online intermediaries are undoubtedly motivated by other concerns, such as facilitating free speech, their ultimate survival as a content or hosting service depends on remaining economically viable.[57] Importantly, a platform's economic viability hinges on its ability to attract new users, sell advertisements, and expand their services to new jurisdictions.[58] Platforms tainted with criminal stigma or weakened by significant fines cannot effectively achieve these goals.[59] Accordingly, rational online intermediaries have decidedly good reason to comply with criminal laws, especially if punishment would severely threaten their expansion capabilities and economic viability.[60]

[54] Commonwealth, *Parliamentary Debates*, House of Representatives, 4 April 2019, 1849–50 (Christian Porter, Attorney-General) (Austl.).

[55] Paternoster, *supra* note 38, at 767–72, 781–82, 784, 787; BECCARIA, *supra* note 38, at 42–43, 93–94; BENTHAM, *supra* note 5, at 19–20, 117, 137.

[56] *See* Byam, *supra* note 10, at 586; Kent Greenfield, *Corporate Constitutional Rights: Easy and Hard Cases*, 98 B.U. L. REV. 40, 41 (2018).

[57] *Free Speech is a Triangle*, *supra* note 6, at 2020.

[58] *Id.* at 2022; Kate Klonick, *The New Governors: The People, Rules, and Processes Governing Online Speech*, 131 HARV. L. REV. 1598, 1625–7 (2018); Danielle Keats Citron & Helen Norton, *Intermediaries and Hate Speech: Fostering Digital Citizenship for Our Information Age*, 91 B.U. L. REV. 1435, 1454 (2011).

[59] *See generally Free Speech is a Triangle, supra* note 6, at 2020 (explaining why infrastructure providers are generally receptive to state pressure).

[60] *See id.*

[14] Given that most content and hosting services undertake rational cost-benefit analyses to maximize profits, it is useful to consider the extent to which the AVM Act adequately incentivizes intermediaries to detect and remove AVM. Under Deterrence Theory, a rational actor will choose to comply with a law if punishment is both sufficiently severe and certain, so the costs of offending outweigh the costs of compliance.[61] When these factors are not present, compliance is reduced significantly.[62] For the AVM Act, this means intermediaries will not be adequately incentivized to remove AVM, which reduces its social utility. The AVM Act seeks to induce compliance by imposing severe criminal penalties on content and hosting services who fail to remove AVM.[63] Corporations may be fined the greater of 10% annual turnover or $11.1 million.[64] Individuals may receive fines of up to $2.22 million, 3 years of imprisonment, or both.[65] Although unclear, it is possible that each subsequent upload of AVM may attract its own penalty, making the potential cost of non-compliance extraordinarily high if the material is continuously re-uploaded by different users.[66]

[61] *E.g.*, Stafford, *supra* note 38, at 255; Paternoster, *supra* note 38, at 769, 782–84; Steven N. Durlauf & Daniel S. Nagin, *Imprisonment and Crime: Can Both be Reduced?*, 10 CRIMINOLOGY & PUB. POL'Y 13, 16 (2011).

[62] *See* Durlauf & Nagin, *supra* note 61, at 17.

[63] *Cf. Abhorrent Materials Act*, *supra* note 1, at s 474.34(10–11) (stating that the penalty for a "body corporate is punishable on conviction by a fine of not more than the greater of the following: (a) 50,000 penalty units; (b) 10% of the annual turnover of the body corporate during the period (the turnover period) of 12 months ending at the end of the month in which the conduct constituting the offence occurred").

[64] *Id.* at s 474.34(10); *Fines and penalties*, AUSTL. SEC. & INV. COMM'N (last updated Sep. 30, 2021), https://asic.gov.au/about-asic/asic-investigations-and-enforcement/fines-and-penalties/ [https://perma.cc/MGN2-NHAJ].

[65] *Abhorrent Materials Act*, *supra* note 1, at s 474.34(9); AUSTL. SECURITIES & INV. COMM'N, *supra* note 64.

[66] Douek, *supra* note 12, at 43.

[15] These severe penalties, and those of similar content removal regimes in other countries, have led Balkin and other authors to conclude that profit-driven content and hosting services will be heavily incentivized to remove illegal content when faced with criminal sanction.[67] Balkin argues that allowing unlawful content to remain online has minimal benefits for social media and hosting companies.[68] These minimal costs are almost always outweighed by the threat of criminal sanctions, which may "hinder [an entity's] ability to do business".[69] Applied to the severe punishment under the AVM Act, a profit-driven intermediary would almost always remove speech it believes to constitute AVM.[70] From an enforcement perspective, this is highly effective.

[16] There is also considerable discourse on the extent to which uncertain enforcement limits deterrence. Indeed, the AVM Act's strict penalties may not induce maximum compliance if the chance of punishment is not sufficiently certain. Criminological research indicates that certainty of punishment is often a better predictor of compliance than severity.[71] Yet, as argued in Section II (A), law enforcement will find it difficult to conduct the monitoring tasks necessary to consistently prosecute failure to remove AVM, particularly in less-publicized cases.[72] This significantly lowers

[67] *E.g.*, *Free Speech is a Triangle*, *supra* note 6, at 2017; Kreimer, *supra* note 7, at 28–29; Felix T. Wu, *Collateral Censorship and the Limits of Intermediary Immunity*, 87 NOTRE DAME L. REV. 293, 300–01 (2011).

[68] *Free Speech is a Triangle*, *supra* note 6, at 2017.

[69] *Id.*

[70] *Abhorrent Materials Act*, *supra* note 1, at s 474.34(10–11).

[71] *See* Paternoster, *supra* note 38, at 817; Daniel S. Nagin & Greg Pogarsky, *Integrating Celerity, Impulsivity, and Extralegal Sanction Threats into a Model of General Deterrence: Theory and Evidence*, 39 CRIMINOLOGY 865, 883–84 (2001).

[72] *See supra* Section II (A).

certainty of punishment, potentially limiting the deterrence induced by the AVM Act's stringent penalties.[73]

[17] In most cases, the threat of extraordinarily strict criminal penalties is still likely to outweigh the benefits of non-compliance, even if there is a lowered threat of prosecution. The major content and hosting platforms will encounter minimal additional compliance costs from removing AVM, already reject AVM under their terms of service, and are intensely profit-driven.[74] Thus, their compliance costs are minimal, but the cost of criminal sanction remains high.[75] However, there are two main groups for which this uncertainty may be fatal for adequate compliance: smaller content and hosting platforms who will struggle to pay compliance costs, and those who are less-profit driven and more committed to hosting AVM. [76]

[18] Contrary to Balkin's claims,[77] monitoring and removing content to the standard required under the AVM Act does not necessarily involve minimal costs, particularly for small intermediaries.[78] The AVM Act's recklessness standard requires platforms to *actively* monitor content to detect AVM.[79] It also requires extremely responsive content moderation procedures to ensure "expeditious" removal, likely to be interpreted by

[73] *See* Paternoster, *supra* note 38, at 817; Nagin & Pogarsky, *supra* note 71, at 883–84; Douek, *supra* note 12, at 57.

[74] *See Free Speech in the Algorithmic Society*, *supra* note 32, at 1179–80, 1182; *see generally* Gillespie, *supra* note 26, at 254 (explaining social media sites' different reasons and methods to remove offensive content).

[75] Douek, *supra* note 12, at 47–48.

[76] *See id.* at 50–53.

[77] *Free Speech is a Triangle*, *supra* note 6, at 2017.

[78] *Id.* at 52–53.

[79] Molly K. Land, *Against Privatized Censorship: Proposals for Responsible Delegation*, 60 VA. J. INT'L L. 363, 384–85 (2020).

courts as within "hours" or "minutes".[80] Monitoring and removing AVM at this intense level requires the implementation of advanced content moderation systems.[81] It will likely involve developing artificial intelligence algorithms to detect questionable material, and hiring moderators to sift through and remove large volumes of content at high speed.[82] While Balkin is correct to assume that removing a particular piece of content is relatively costless, the detection procedures required to facilitate removal under the AVM Act are extremely expensive.[83] Larger intermediaries such as Facebook and Twitter will be highly capable of absorbing this compliance cost, however smaller intermediaries may struggle. Combined with a low certainty of punishment, a profit-driven small content or hosting service may rationalize non-compliance and fail to implement the moderation systems required to adequately monitor and remove AVM on their platforms, a decidedly undesirable outcome for enforcement.

[19] Additionally, some content and hosting services appear to be less profit-driven and more committed to offending, so uncertain strict punishment may not adequately incentivize compliance.[84] Broader criminological evidence indicates individuals who are committed to

[80] Douek, *supra* note 12, at 45.

[81] *See id.* at 52.

[82] *See* Gillespie, *supra* note 26, at 266–67 (human moderators); Bloch-Wehba, *supra* note 6, at 56–57 (artificial intelligence).

[83] *Compare Free Speech is a Triangle*, *supra* note 6, at 2017, *with* Douek, *supra* note 12, at 52–53 (noting that the resources required for such detection procedures are not generally as readily available for small businesses as they are for the major platforms); Bloch-Wehba, *supra* note 6, at 91 (noting elevated compliance costs for small intermediaries).

[84] *See* Jeff A. Bouffard et al., *Examining the Stability and Predictors of Deterrability Across Multiple Offense Types Within a Sample of Convicted Felons*, 57 J. CRIM. JUST. 76, 77 (2018); Greg Pogarsky, *Identifying "Deterrable" Offenders: Implications for Research on Deterrence*, 19 JUST. Q. 431, 433 (2002).

offending are often unresponsive to uncertain punishment.[85] One notable example in the content-moderation context is 8chan, a social media site particularly tolerant of violent content.[86] If the AVM Act can be successfully enforced, the potential fines will be so large that they could eventually force these platforms to cease operation.[87] If punishment remains uncertain, the potential deterrent effect will be limited, as these platforms will have no identifiable reason to change their stance on supporting AVM.[88]

[20] Therefore, it is suggested that the AVM Act's enforcement inefficiencies and uncertainty undermine the otherwise strong incentive for platforms to remove AVM. Hence, the AVM Act fails to fully capitalize on the expected enforcement benefits of public-private collaboration.[89] In most cases, the extremely strict penalties clearly outweigh the cost of compliance.[90] However, sub-optimal certainty of punishment may viably cause an intermediary struggling with compliance costs, or with bad intentions, to take the risk and fail to implement appropriate content moderation systems. Accordingly, the AVM Act does not effectively enforce the removal of AVM, nor does it maximize social utility to the anticipated extent.

III. The AVM Act and Freedom of Speech

[21] Beyond considerations of enforcement mechanisms, examining the merits of the AVM Act must include an analysis of its potential to induce

[85] Bouffard et al., *supra* note 84, at 77; Pogarsky, *supra* note 84, at 433.

[86] *See* Douek, *supra* note 12, at 51.

[87] *Cf. Free Speech is a Triangle*, *supra* note 6, at 2017.

[88] *See generally* Bouffard et al., *supra* note 84, at 77 (discussing the positive relationship between certainty of punishment and deterrence); Pogarsky, *supra* note 84, at 433 (explaining how the threat of punishment is ineffective in deterring certain offenders).

[89] *See* Douek, *supra* note 12, at 53.

[90] *Id.*

collateral over-removal of lawful content on social media platforms. Under the UPC, any benefits produced by a criminal law must not be outweighed by any collateral harm the law creates.[91] However, the AVM Act contravenes this principle by inducing collateral over-removal of beneficial speech, beyond what is necessary to achieve a reduction in AVM. As shown in Section II, the AVM Act does not adequately incentivize all social media and hosting companies to remove AVM because punishment is uncertain.[92] Building on the understanding of social media companies' rational, profit-driven nature discussed in Section II, this Section argues that, even when the AVM Act *does* induce compliance, it does so at the expense of highly socially valuable, non-AVM content. This trade-off is unconducive to maximizing social utility and fulfilling the UPC, as similar enforcement outcomes could be achieved with reduced burdens on beneficial speech.

A. The AVM Act and the Potential Over-Removal of Legitimate Speech

[22] Determining whether a particular piece of content violates the AVM Act is highly dependent on contextual factors.[93] Violent conduct filmed by a concerned bystander does not fall within the AVM Act's ambit, yet a near identical video filmed by the perpetrator is a clear violation.[94] Even content that would otherwise constitute AVM may attract a defense if it falls within one of several permitted uses.[95] These permitted uses include: enforcing monitoring or investigating a breach of Australian law; court or tribunal proceedings; research purposes; professional public interest journalism; a public official's duties or functions; advocating for "the lawful procurement of a change" to Australian "law, policy or practice", or; "development,

[91] Søbirk Petersen, *supra* note 5; *see also* Szott Moohr, *supra* note 5, at 786.

[92] *See supra* pp. 14–15.

[93] Bloch-Wehba, *supra* note 6, at 45, 70, 88; Douek, *supra* note 12, at 48.

[94] *Abhorrent Materials Act*, *supra* note 1, at s 474.31(c).

[95] *Id.* at s 474.37.

performance, exhibition or distribution" of an artistic work.[96] Additionally, the AVM Act does not apply to political content.[97] These intricacies will be clear to a court determining whether a social media or hosting platform has violated the AVM Act. However, the difference between unlawful AVM and legitimate content is not always clear to the platforms, who are unskilled in legal analysis, attempting to detect and "expeditiously" remove content in practice.[98]

[23] Concerningly, social media and hosting services' content moderation systems cannot easily detect these contextual nuances, particularly when operating under the AVM Act's strict content removal timeframes. The AVM Act's stringent detection and removal requirements implicitly require platforms to utilize moderation algorithms to detect and remove content.[99] However, content moderation algorithms do not currently have the technical capacity to distinguish between a violent act filmed by a perpetrator, and a similar video filmed by a concerned bystander.[100] They also have trouble identifying genuine AVM when the original content has

[96] *Id.* at s 474.37(1)–(2).

[97] *Id.* at s 474.38.

[98] *See* Douek, *supra* note 12, at 45–46; Bloch-Wehba, *supra* note 6, at 62–63; *see generally* Annemarie Bridy, *Remediating Social Media: A Layer-Conscious Approach*, 24 B.U. J. SCI. & TECH. L. 193, 226 (2018) (discussing how self-driven removals may be more accurate than computer software techniques).

[99] Bloch-Wehba, *supra* note 6, at 45–46.

[100] *See, e.g.*, Douek, *supra* note 12, at 48; Bloch-Wehba, *supra* note 6, at 77–78; Bridy, *supra* note 98, at 219.

been modified.[101] For example, Facebook's moderation systems could not detect edited versions of the Christchurch attack video.[102] Moreover, even if a moderation algorithm correctly recognizes genuine AVM, it will be incapable of determining whether the content falls within one of the permitted exceptions. To an algorithm, AVM used for nefarious purposes looks identical to the same video used for genuine news reporting, artistic, or political purposes.[103] Accordingly, it is currently impossible to detect and remove genuine AVM to the standard required under the AVM Act without also removing legitimate content. In practice, content moderation systems are likely to falsely detect legitimate content in some cases, while simultaneously failing to identify genuine AVM in others,[104] rendering current moderation capabilities incongruent with the AVM Act's requirements. While a human moderator may be better equipped to identify these contextual nuances and make accurate removal decisions,[105] given the extremely large volume of content uploaded to social media every second,

[101] Bloch-Wehba, *supra* note 6, at 56; *see also* Janis Dalins et al., *PDQ & TMK + PDQF - A Test Drive of Fakebook's Perceptual Hashing Algorithms*, J. DIGIT. INVESTIGATION, Dec. 18, 2019, at 5–7 (studying which algorithms are less effective in efforts to remove AVM materials).

[102] Douek, *supra* note 12, at 51; Peter A. Thompson, *Beware of Geeks Bearing Gifts: Assessing the Regulatory Response to the Christchurch Call*, 7 POL. ECON. COMM'N 83 (2019); Gavin Ellis and Denis Muller, *The Proximity Filter: The Effect of Distance on Media Coverage of the Christchurch Mosque Attacks*, 15 KOTUITUI: N.Z. J. SOC. SCI. ONLINE 332, 335 (2020).

[103] *See* Bloch-Wehba, *supra* note 6, at 76; Stuart Macdonald et al., *Regulating Terrorist Content on Social Media: Automation and the Rule of Law*, 15 INT'L. J.L. CONTEXT 183, 190 (2019); *cf.* Bridy, *supra* note 98, at 226–27 (explaining, via an example from YouTube's system, how human self-performed removals may be more accurate than computer-based ones).

[104] Douek, *supra* note 12, at 51; Bloch-Wehba, *supra* note 6, at 77–78.

[105] *See* Macdonald et al., *supra* note 103; Bridy, *supra* note 98, at 226.

human moderators cannot examine each piece of content to determine its legality within the short timeframe required.[106] Coupled with the fact that platforms must remove AVM "expeditiously" under threat of extreme criminal sanctions, even human moderators will have difficulty making these complex decisions.[107]

[24] In light of such limitations in current moderation capabilities, a social media platform or hosting service is faced with two options. One option is to develop more *inclusive* algorithms systems to detect AVM, removing all content caught by these systems, regardless of whether the content actually constitutes AVM. This would result in increased AVM removal and avoidance of criminal sanction, but at the expense of removing legitimate, non-AVM content.[108] Conversely, the other option is to direct these systems and human moderators to be less stringent in their removal, almost certainly missing actual AVM and risking extremely severe criminal punishment, but preserving more legitimate non-AVM content.[109] In such a context, it is foreseeable that some platforms will forgo compliance altogether due to uncertainty of punishment.[110] In these cases, the AVM Act produces no substantial effect on either AVM or legitimate content and produces no additional utility.[111] However, for those platforms who wish to cooperate with the AVM Act to avoid any possibility of strict criminal

[106] *See* Macdonald et al., *supra* note 103, at 184; *see generally* Céline Castets-Renard, *Algorithmic Content Moderation on Social Media in EU Law: Illusion of Perfect Enforcement*, 2020 U. ILL. J.L. TECH & POL'Y 283, 293 (2020) (recommending mechanisms for making algorithmic systems more transparent).

[107] *See* Douek, *supra* note 12, at 48; *see generally* Castets-Renard, *supra* note 106, at 310–11 (discussing the legal duty that online platforms have to remove AVM).

[108] *See* Douek, *supra* note 12, at 48–49; *Free Speech is a Triangle*, *supra* note 6, at 2017, 2019; Bloch-Wehba, *supra* note 6 at 75, 78.

[109] *See, e.g.*, Douek, *supra* note 12, at 51; *see generally Free Speech is a Triangle*, *supra* note 6, at 2018 (explaining the problems with being overly-inclusive in content removal).

[110] *See supra* pp. 15–17.

[111] *See supra* pp. 18–19.

sanction or stigma, they are effectively forced to choose the first option.[112] In this way, compliance may necessitate over-removal.

[25] While it is conceivable that technologies could be developed to create more accurate algorithms that can filter AVM and also minimize the inclusion of legitimate non-AVM content,[113] the AVM Act does not incentivize the development of such systems.[114] Instead, it induces the opposite.[115] By criminalizing the failure to remove AVM and imposing significant criminal sanctions, the AVM Act incentivizes those who choose to comply to block *more* content for fear of criminal liability and stigma.[116] Extensive empirical evidence supports this claim.[117] For example, Urban and Quilter's study of copyright liability regimes found that platforms complied with removal requests regardless of whether the content actually infringed copyright.[118] Although not specific to the AVM Act, this evidence strongly suggests moderation systems will become more *inclusive*, not more accurate.[119]

[112] *See* Douek, *supra* note 12, at 47–49; *Free Speech is a Triangle*, *supra* note 6, at 2017, 2019–20.

[113] *See generally* Bloch-Wehba, *supra* note 6, at 41–42 (describing the technical capabilities of AutoModerators).

[114] *Id.* at 88.

[115] *Id.* at 74–77.

[116] *Guarding the Guardians*, *supra* note 6, at 678.

[117] *See* Sharon Bar-Ziv & Niva Elkin-Koren, *Behind the Scenes of Online Copyright Enforcement: Empirical Evidence on Notice & Takedown*, 50 CONN. L. REV. 339, 345 (2018); *see also* Jennifer M. Urban & Laura Quilter, *Efficient Process or Chilling Effects - Takedown Notices Under Section 512 of the Digital Millennium Copyright Act*, 22 SANTA CLARA HIGH TECH. L.J. 621, 681 (2006); Jennifer Urban et al., *Notice and Takedown: Online Service Provider and Rightsholder Accounts of Everyday Practice*, 64 J. COPYRIGHT SOC'Y U.S.A. 371, 403, 407 (2017).

[118] Urban et. al., *supra* note 117, at 626; Land, *supra* note 79, at 411–12.

[119] *See Guarding the Guardians, supra* note 6, at 678; Bloch-Wehba, *supra* note 6, at 65.

B. The AVM Act and Socially-Beneficial Speech

[26] Although compliance with the AVM Act is likely to induce collateral over-removal of legitimate content, this alone is insufficient to establish that the AVM Act is inadequate. Under the UPC, it may be argued that over-removal is justified, if it is necessary to ensure AVM is removed and produces a net social benefit.[120] Australia does not have a positive right to freedom of speech, and the AVM Act does not apply to the extent that it would encroach on the implied freedom to political communication.[121] Thus, there is no *legal* reason as to why this collateral impact is concerning.[122] Nevertheless, the types of legitimate speech likely to be removed by social media companies attempting to adhere to the AVM Act provide varying degrees of social utility.[123] Thus, there are strong utilitarian reasons to doubt the AVM Act's adequacy.

[27] There are two main categories of legitimate content that social media and hosting companies may unjustifiably remove in their attempts to implement the AVM Act. The first category includes violent conduct recorded by an activist or concerned bystander, such as videos of police brutality and war crimes. [124] The second category includes AVM that would otherwise fall under an exception, such as online journalistic news reports containing AVM. [125]

[120] *See* Land, *supra* note 79, at 416 (discussing how over-removal may be justified if it is necessary and produces a net social benefit).

[121] *See generally Lange v Austl. Broad. Corp.* (1997) 189 CLR 520 (Austl.) (discussing how the freedom of communication is not a privilege); *Abhorrent Materials Act*, *supra* note 1, at s 474.34.

[122] *See also* Land, *supra* note 79, at 414–15 (noting that there may be international law concerns associated with privatized censorship regimes, such as the AVM Act).

[123] Douek, *supra* note 12, at 43.

[124] *See id.* at 56.

[125] *Id.* at 48, 53.

[28] Under utilitarian and consequentialist justifications for freedom of speech, a particular type of speech is justified if it produces a social benefit.[126] Mill's utilitarian justification argues that free speech is beneficial because it facilitates knowledge of the truth, which itself produces utility by contributing to the "marketplace of ideas."[127] Aside from truth, other consequentialist reasons highlight free speech's utility in promoting democracy and holding governments accountable.[128] However, the adequacy of these justifications depends on whether the type of speech in question is actually conducive to achieving a beneficial end result.[129] Indeed, many of these theories have been disputed as a broad justification for a positive right to free expression, particularly because not all forms of expression are necessarily conducive to achieving any sort of beneficial consequence.[130] Examples include child pornography or genuine AVM.[131] Under the UPC, removing these forms of content is more readily justified to produce an overall public benefit. Yet unlike low-value speech, these utilitarian justifications for speech are particularly applicable to the types of high-value, legitimate content likely to be wrongly removed under the AVM Act.

[126] HARRY MELKONIAN, FREEDOM OF SPEECH AND SOCIETY: A SOCIAL APPROACH TO FREEDOM OF EXPRESSION 98 (Cambria Press 2012).

[127] JOHN STUART MILL, ON LIBERTY 35–37 (The Walter Scott Publishing Co. 2011) (1859); WOJCIECH SADURSKI, FREEDOM OF SPEECH AND ITS LIMITS 8 (Kluwer Academic Publishers 1999); MELKONIAN, *supra* note 126, at 101–06.

[128] Christopher T. Wonnell, *Truth and the Marketplace of Ideas*, 19 U.C. DAVIS L. REV. 669, 669–70 (1980) (discussing the utility of free speech in promoting democracy and holding governments accountable); Vincent Blasi, *The Checking Value in First Amendment Theory*, 3 AMERICAN B. FOUND. RES. J. 521, 524–27 (1977); MELKONIAN, *supra* note 126 at 115–16.

[129] Twana A. Hassan, *Critiques of the Pursuit of Truth as a Justificatory Theory of Free Speech*. 3 INT'L. J. HUM. RTS. 171, 171–79 (2015).

[130] MELKONIAN, *supra* note 126, at 107–08; Hassan, *supra* note 129 at 171, 179.

[131] Bloch-Wehba, *supra* note 6, at 58.

[29] Unlike low-value forms of speech, or speech based on mere opinion, the very purpose of a concerned bystander or activist recording and uploading violent material is to directly expose true events.[132] Depending on the exact content involved, exposing these truths through social media can be highly beneficial for democratic or political reasons.[133] If the violence involves a state crime, such as police brutality, activists use this content to hold leaders accountable for atrocities that would otherwise be unknown.[134] In turn, this enables citizens to make more informed political decisions based on the reality of their leaders' actions.[135]

[30] A key example occurred during the "Arab Spring" protests of 2011 and 2012.[136] During these protests, activists recorded videos of extreme government violence against citizens and uploaded them to YouTube.[137] This resulted in "mobilising . . . public and global opinion against the atrocities," producing significant social utility.[138] Given these significant social benefits, a law that unnecessarily impinges upon this type of content considerably lowers overall utility.

[31] The types of AVM that likely fall under a defense also provide varying degrees of social benefit, albeit some clearer than others. For

[132] Douek, *supra* note 12, at 48.

[133] Michael Karanicolas, *Understanding the Internet as a Human Right*, 10 CAN. J.L. & TECH. 263, 265–67 (2012).

[134] Douek, *supra* note 12, at 48.

[135] *See generally* MELKONIAN, *supra* note 126, at 115–16 (citing ALEXANDER MEIKLEJOHN, FREE SPEECH AND ITS RELATION TO SELF-GOVERNMENT (Harper & Brothers Publishers 1948) (explaining Alexander Meiklejohn's free speech theory that describes how broad elements of free speech directly and indirectly affects self-government)).

[136] Karanicolas, *supra* note 133, at 266.

[137] *Id.* at 267.

[138] *Id.*

example, AVM used in public interest journalism facilitates truth discovery, and preventing journalistic access to these materials would severely hinder online news reporting.[139] Further, AVM used to investigate breaches of the law, or used in court proceedings, facilitates Australia's vital legal processes.[140] Without access to these materials, law enforcement would struggle to access the evidence necessary to investigate and prosecute violent crimes, which is clearly undesirable. Even AVM used for artistic purposes, which may appear less justifiable, provides some benefit.[141] It may provide entertainment or cultural benefit, or express important political messages to a worldwide audience.[142] For example, He Yunchang's performance artwork, "One Meter Democracy," depicts the artist receiving a graphic "meter-long incision down the length of his body without anesthetics [sic]."[143] Although violent, the artwork is a political commentary on the "tension between the individual and the state" that seeks to induce deliberation in a viewer's mind and inform their political decision-making.[144] All of these socially-beneficial content categories would be at risk under the AVM Act, reducing its overall utility.

[32] Critics may argue that this socially-beneficial content could simply be accessed or disseminated through other methods aside from social media,

[139] *See* Douek, *supra* note 12, at 53.

[140] Explanatory Memorandum, *supra* note 4, at paras 31, 109–11 (Austl.).

[141] MELKONIAN, *supra* note 126, at 116 (citing Alexander Meiklejohn, *The First Amendment is an Absolute*, 1961 SUP. CT. REV. 245, 256–57 (1961)); *see Campbell v MGN Ltd.* [2004] 2 AC 457, 499–500 (Austl.).

[142] *See* MELKONIAN, *supra* note 126, at 1.

[143] Jing Cao, *He Yunchang: Water Forming Stone at Ink Studio*, DAILYSERVING (Jan. 27, 2016), https://web.archive.org/web/20170322034619/https://www.dailyserving.com/2016/01/he-yunchang-water-forming-stone-at-ink-studio/ [https://perma.cc/5GLW-6C75].

[144] Andrea Mejía, *The Artist Who Tortures Himself to Create Masterpieces of Pain*, CULTURA COLECTIVA (Oct. 6, 2017), https://culturacolectiva.com/art/he-yunchang-pain-performances [https://perma.cc/2SDW-5WEL].

so the impact of over-removal is exaggerated. Indeed, an activist attempting to expose a war crime may be able to smuggle the video to journalists through traditional means, a court or researcher may still be able to access AVM without social media, and an artist may still display their work in a physical gallery.[145] Thus, this criticism has some degree of merit.

[33] However, traditional methods of obtaining and disseminating this type of material do not *maximize* utility to the same degree as social media. The reach and accessibility of social media is unprecedented, which amplifies the existing utility of this socially-beneficial content.[146] It is now one of the "principal sources for knowing current events," and has been referred to as "the modern public square".[147] Unlike traditional media, it is accessible anywhere in the world, to anyone with an internet connection.[148] Thus journalists, artists, and courts can publish socially-beneficial content more *quickly*, to a *broader* audience, regardless of their social standing.[149] Restricting this content to traditional media would fail to capitalize on the additional utility social media offers and deprive the "modern public square" of highly beneficial content.[150] A law that impinges upon this public benefit without justifiable reason, even if it does so unintentionally, significantly lowers social utility.

[145] *See* Karanicolas, *supra* note 133, at 267.

[146] *See e.g.*, *id.* at 266–67.

[147] Packingham v. North Carolina, 137 S. Ct. 1730, 1732 (2017).

[148] Aaron D. White, *Crossing the Electronic Border: Free Speech Protection for the International Internet*, 58 DEPAUL L. REV. 491, 491–92 (2019).

[149] *See, e.g.*, Karanicolas, *supra* note 133, at 267.

[150] *Packingham*, 137 S. Ct. at 1732.

C. Calibrating the Removal of AVM with the Maintenance of Free Speech

[34] While atrocities such as the Christchurch attack are extremely socially damaging so as to justify *some* collateral restrictions on socially-beneficial speech, the AVM Act sacrifices more valuable speech than is necessary to mitigate this harm. A similar level of enforcement can be achieved with fewer collateral impacts by simply introducing accountability safeguards, such as mandatory appeals processes and transparency reporting similar to those in other jurisdictions.[151]

[35] Requiring social media companies to implement a mandatory appeals mechanism will provide users an opportunity to appeal incorrect removal decisions and have their content reinstated.[152] Additionally, mandatory transparency reporting requirements enable users to know how and when their content is removed, facilitating better public oversight.[153] While these measures are imperfect and leave the AVM Act's deeper issues unaddressed (explored in detail in Section IV), they produce greater overall utility than the current AVM Act and better align with the UPC.

[36] While the AVM Act's potential enforcement benefits are lower than legislators anticipated, the negative effect on public interest speech is higher than originally thought. By failing to recognize that compliance necessitates over-inclusiveness and refusing to implement safeguards, the AVM Act's legislators implicitly chose to value *increased* content removal over *accurate* content removal. However, this does not result in maximum social

[151] *See NetzDG*, *supra* note 43; FULL GOVERNMENT RESPONSE, *supra* note 43.

[152] *See* Bloch-Wehba, *supra* note 6, at 90; Directive 2019/790, of the European Parliament and of the Council of 17 April 2019 on Copyright and Related Rights in the Digital Single Market and Amending Directives 96/9/EC and 2001/29/EC, art. 17(9), 2019 O.J. (L 130) 92 (EU) [hereinafter EU Copyright Directive]; Bridy, *supra* note 98, at 226–27, n. 200.

[153] Douek, *supra* note 12, at 52, 59; *see generally* Bloch-Wehba, *supra* note 6, at 87-8; Keats-Citron, *supra* note 15.

utility, nor does it uphold the UPC.[154] It places an unnecessarily harsh burden on socially-beneficial speech, without generating any additional enforcement benefits.[155] A more measured approach, including safeguards for public-interest speech, would better maximize overall utility.

IV. Relevant International Developments

[37] Australia's enactment of the AVM Act has led to it being a world leader in the regulation of AVM. Indeed, then, it is useful to situate the Australian legislative framework within broader international developments by considering the United Kingdom's *Online Safety Bill*.[156] The United Kingdom *Government Response to the Online Harms White Paper,* presented to Parliament in December 2020, announced a new *Online Safety Bill* to reduce the sharing of harmful content on the internet.[157] However, the United Kingdom Bill will adopt a very different approach to online safety to that of the Australian AVM Act. While the centerpiece of the Australian legislation is the imposition of a statutory obligation to remove AVM, the United Kingdom Bill will impose a wider new statutory duty of care on online entities to take appropriate responsibility for the safety of the users of their platforms and services.[158] For purposes of the present discussion, it is relevant to note that this new statutory duty will include taking responsibility for "user generated content."[159] "User generated content" will be defined as "digital content (including text,

[154] *See* Søbirk-Petersen, *supra* note 5, at 126; Szott Moohr, *supra* note 5, at 786; BENTHAM, *supra* note 5, at 134–35.

[155] *See* Douek, *supra* note 12, at 53.

[156] DEPARTMENT FOR DIGITAL, CULTURE, MEDIA & SPORT, ONLINE SAFETY BILL DRAFT, 2021, Cm. 405 (UK) [hereinafter ONLINE SAFETY BILL DRAFT].

[157] FULL GOVERNMENT RESPONSE, *supra* note 43, at 13.

[158] *Abhorrent Materials Act*, *supra* note 1, s 474.34; FULL GOVERNMENT RESPONSE, *supra* note 43, at 16.

[159] FULL GOVERNMENT RESPONSE, *supra* note 43, at 16.

images and audio) produced, promoted, generated or shared by users of an online service," and "content may be paid-for or free, time-limited or permanent."[160] "User" will be defined to be "any individual, business or organisation (private or public) that puts content on a third-party online service."[161]

[38] Significantly, the *Online Safety Bill*'s definition of "harmful content," which will attract this new statutory duty of care, will be broader than the definition of "abhorrent violent material" in the Australian AVM Act.[162] "Harmful content" in the United Kingdom Bill will be defined as content or activity which "gives rise to a reasonably foreseeable risk of harm to individuals, and which has a significant impact on users or others".[163] While the new Bill will principally address online illegal activity by seeking to prevent children from being exposed to inappropriate material, it will also address other types of harmful online content, including most notably disinformation and misinformation that is not specifically directed at children.[164] Examples provided in the *Government Response* include inaccurate information as to the safety of vaccines and destructive pro-anorexia content.[165] By contrast, "abhorrent violent material" is fairly narrowly defined in the Australian legislation to be audio or video recorded by a perpetrator or accomplice, depicting terrorism, rape, torture, murder, attempted murder or kidnapping.[166]

[160] *Id.*

[161] *Id.*

[162] *Compare* FULL GOVERNMENT RESPONSE, *supra* note 43, at 23–51, *with Abhorrent Materials Act*, *supra* note 1, at ss 474.31, 474.32(1).

[163] FULL GOVERNMENT RESPONSE, *supra* note 43, at 23.

[164] *Id.* at 3–4.

[165] *Id.* at 4.

[166] *Abhorrent Materials Act*, *supra* note 1, at ss 474.31, 474.32(1).

[39] In order to satisfy the new statutory duty of care, regulated entities in the United Kingdom will be required to enact appropriate systems and processes to strengthen online safety in line with Codes of Practice developed by the Office of Communications, commonly known as Ofcom.[167] The Bill will introduce differentiated expectations for online entities assessed on the basis of the nature of the content and activity on their services.[168] The Bill will establish three categories of content – content which is illegal, content which is harmful to children, and content which is legally accessible by adults but which nonetheless may be harmful.[169] It will hence adopt a tiered approach with different regulatory requirements imposed on different services.[170] It is relevant to note that the new Bill will not alter liability of online companies for illegal content that satisfies the relevant definition of harmful content. Rather it requires regulated entities to enact "appropriate systems and processes" to protect their users.[171] Only a high-risk category services will be required to take steps address harmful content.[172] The *Government Response* emphasizes that duty of care will be directed at systems and processes rather than individual pieces of content.

[40] Perhaps most significantly, the proposed *Online Safety Bill* differs from the AVM Act in its approach to the governance of internet service providers (ISPs). The United Kingdom Bill will apply to entities whose services either host user-generated content that can be accessed by users in the UK, and/or those that facilitate public or private online interaction between service users, one or more of whom is in the United Kingdom.[173]

[167] FULL GOVERNMENT RESPONSE, *supra* note 43, at 31–32, 54.

[168] *Id.* at 10.

[169] *Id.*

[170] *Id.*

[171] *Id.* at 5.

[172] *Id.* at 29.

[173] *Id.* at 15.

However, the *Online Safety Bill* will not apply to entities who fulfil merely "a functional role in enabling online activity."[174] Interestingly, ISPs are included within the definition of entities playing merely a functional role and will be exempt from the duty of care. [175] This differs from the AVM Act which stipulates that ISPs may be liable if they are aware of content depicting abhorrent violent conduct (that has occurred or is taking place in Australia) being made accessible through their services, yet fail to refer the content to the Australian Federal Police.[176] Interestingly, although the internet intermediaries of ISPs are exempt, the Bill will apply to search engines, another species of internet intermediaries.[177] The *Government's Response* justifies this distinction on the basis that while search engines do not directly host user-generated content, they facilitate easy access to harmful content.[178] This somewhat fragmented application of the United Kingdom Bill differs from the comprehensive coverage of the Australian AVM Act considered above.

[41] However, similar to the Australian AVM Act, the United Kingdom Bill will seek to calibrate online safety and freedom of speech. The United Kingdom Bill will do so by limiting its application to entities who are considered to pose a substantive risk to online safety.[179] It will apply to companies whose service hosts user-generated content that can be accessed by users in the UK, as well as those that facilitate public or private online interaction between service users, one or more of whom is in the UK.[180]

[174] *Id.*

[175] *See id.* at 9 (specifying that regulated entities do however have a duty to cooperate with the regulator on the enactment of appropriate business disruption measures).

[176] *See Abhorrent Materials Act supra* note 1, at s 474.33.

[177] FULL GOVERNMENT RESPONSE, *supra* note 43, at 9.

[178] *Id*. at 17.

[179] *Id*. at 3–4.

[180] *Id.* at 16.

However, certain low-risk businesses, such as retailers who offer only product and service reviews, will be exempt from the duty of care.[181] Supporting the above measures to advance online safety are a variety of provisions designed to uphold freedom of expression and media freedom. The proposed legislation will require online entities to maintain accessible and effective complaint mechanisms to enable users to object to unfair removal of content.[182] Further, the Bill will also provide a spate of specific protections for journalistic content shared on in-scope services.[183] This express and detailed commitment to protecting freedom of speech is useful for Australian law and policy makers seeking to refine Australia's AVM Act so that it better aligns the creation of a safe online environment with the maintenance of the right to freedom of expression. The next section of this paper considers this issue, along with further legislative reforms aimed at enhancing the effectiveness of the Australian AVM Act.

V. Options for Reform and Refinement

[42] On the basis of the above critique, it is suggested that the AVM Act possesses two main areas that require amendment to maximize its overall utility. Criminalizing the failure to remove AVM, without implementing any transparency requirements, makes transgressions highly difficult to detect. This diminishes certainty of punishment and fails to induce maximum removal of AVM. Moreover, the AVM Act encourages over-removal of socially-beneficial speech without implementing any safeguards.[184]

[43] Nevertheless, there are some aspects of the current AVM Act that should remain unaltered. The AVM Act's utilization of public-private cooperation by targeting content and hosting services rather than individual

[181] *Id.* at 9.

[182] *Id.* at 4.

[183] Full Government Response, *supra* note 43, at 15.

[184] *Cf. NetzDG*, *supra* note 43; Full Government Response, *supra* note 43.

users, is crucial for enforcement and should remain in any future amendment.[185] The AVM Act also rightly recognizes the rational, profit-driven nature of most social media platforms by imposing strict criminal sanctions for those who fail to comply.[186] Combined, these factors are conducive to a reduction in AVM, and increase the AVM Act's social utility.

[44] In this regard, one option would be to amend the AVM Act to introduce mandatory transparency reporting requirements and appeals mechanisms. However, these measures would not wholly address the AVM Act's underlying issues. Criminalizing the failure to remove AVM is not conducive to efficient enforcement, or minimizing over-removal of legitimate speech. Thus, to maximize overall utility in the long term, the AVM Act's main offense should be amended to criminalize a failure to implement appropriate content moderation mechanisms.[187] This Section also recommends the establishment of a regulatory body to oversee this new offense and determine what an appropriate moderation system should entail.[188] While this approach is more wide-reaching than the current AVM Act, it is arguably necessary to produce maximum social benefit and better-align with the UPC.

[185] *See* Lichtman & Posner, *supra* note 29, at 221, 233; Peterson et al., *supra* note 30, at 598; *Governance by Proxy*, *supra* note 9, at 113–15.

[186] *See* Paternoster & Simpson, *supra* note 10, at 579–80; Sanford H Kadish, *Some Observations on the Use of Criminal Sanctions in the Enforcement of Economic Regulations*, in WHITE COLLAR CRIME: OFFENSES IN BUSINESS, POLITICS AND THE PROFESSIONS 426 (G Geis & R.F. Meier eds., 1977).

[187] *See* REPUBLIC OF FR., CRÉER UN CADRE FRANÇAIS DE RESPONSABILIZATION DES RÉSEAUX SOCIAUX: AGIR EN FRANCE AVEC UNE AMBITION EUROPÉENNE [CREATING A FRENCH FRAMEWORK TO MAKE SOCIAL MEDIA PLATFORMS MORE ACCOUNTABLE: ACTING IN FRANCE WITH A EUROPEAN VISION] 1, 10 (2019).

[188] FULL GOVERNMENT RESPONSE, *supra* note 43, at 56.

A. Minimum Measures to Maximize Enforcement and Minimize Legitimate Speech Removal

[45] Mandatory transparency reporting is often cited as the core requirement of a law seeking to regulate social media content, making it an important amendment to remedy the AVM Act's deficiencies.[189] As discussed in Section II, compliance cannot be adequately monitored without an understanding of social media companies' content removal practices.[190] If social media and hosting companies are required to report on their efforts to comply with the AVM Act and their content removal practices, law enforcement can focus its attention on platforms that are forgoing compliance.[191] Ideally, this will assist in the detection of content removal failures, increase certainty of punishment, and by extension, encourage the removal of AVM.[192] Aside from assisting law enforcement, transparency reporting will also ensure users are better informed about how and when their content is removed.[193] Although transparency alone will not solve the over-removal issue, it provides a foundation for users to challenge incorrect removal decisions and critique platforms' broader content moderation processes.[194] Thus, mandatory transparency reporting should be seen as a crucial baseline reform to facilitate enforcement and lay the foundations to reduce over-removal. As Suzor notes, generic calls for "transparency" in content moderation are largely unhelpful in accurately determining how

[189] *See, e.g.*, Nicholas P. Suzor et al., *What Do We Mean When We Talk About Transparency? Toward Meaningful Transparency in Commercial Content Moderation*, 13 INT'L. J. COMMC'N. 1526, 1527–28 (2019).

[190] *See supra* Section II.

[191] *See* FULL GOVERNMENT RESPONSE, *supra* note 43, at 58.

[192] *See* Kadish, *supra* note 186, at 426; *cf.* Stafford, *supra* note 38, at 255–56 (explaining how deterrence is the "omission or curtailment of crime out of fear of legal punishment.").

[193] Keats-Citron, *supra* note 15, at 1067–68.

[194] *See generally* Suzor et al., *supra* note 189, at 1529.

social media and hosting services are regulating their content.[195] Thus, it is necessary to determine the specific *type* of transparency reporting that the AVM Act should mandate to best ameliorate its deficiencies. [196]

[46] Content moderation regimes in other jurisdictions provide some guidance on what a potential transparency requirement might involve. Germany's *NetzDG* mandates that social media companies submit bi-annual reports detailing how they responded to specific unlawful content complaints.[197] Conversely, the UK's *Online Harms White Paper* empowers an independent regulator to request reports on companies' efforts to comply with law enforcement, and the tools and processes they use to remove unlawful content on their platforms.[198] A similar requirement introduced to the AVM Act could require social media companies to detail their efforts to remove AVM, and disclose statistical data detailing the amount and type of lawful and unlawful content removed. This information would be useful to highlight potential enforcement deficiencies, and identify the types of legitimate content that are being unjustifiably caught by moderation systems.[199]

[47] Nevertheless, the AVM Act requires more comprehensive transparency measures. Due to the AVM Act's stringent removal requirements, many social media companies will rely on *ex ante* artificial intelligence moderation algorithms that filter AVM *before* it is uploaded.[200] Thus, simply disclosing information and data about content that was

[195] *Id. (*noting that generic calls for "transparency" in content moderation are largely unhelpful to accurately determine how social media and hosting services are regulating their content).

[196] *Id.*

[197] *NetzDG*, *supra* note 43.

[198] DEPARTMENT FOR DIGITAL, CULTURE, MEDIA AND SPORT, ONLINE HARMS WHITE PAPER, 2019, Cm. 57, at 9 (UK) [hereinafter ONLINE HARMS WHITE PAPER].

[199] *See generally* Suzor et al., *supra* note 189, at 1529; Douek, *supra* note 12, at 52.

[200] Bloch-Wehba, *supra* note 6, at 53, 55.

removed *ex post* will not provide a complete picture of the underlying systems that make these filtering and removal decisions in the first instance.[201] Accordingly, the AVM Act should not only mandate transparency into the amount and type of content removed, but also the moderation algorithms that filter AVM.[202] This algorithmic transparency requirement will provide a more complete picture of how platforms are moderating their content in order to identify potential issues more accurately.

[48] While mandatory transparency reporting is a valuable facet of reform, it cannot address all of the problems associated with the AVM Act. Artificial intelligence-driven content moderation algorithms constantly self-learn, updating in response to new content uploaded to the platform.[203] Such algorithms use information about existing content to make moderation decisions about content uploaded in the future, and then update their processes accordingly.[204] Moreover, even the algorithm's creators often cannot anticipate how it will develop in practice.[205] This constant self-learning means a transparency report may only be accurate for a short time following disclosure, limiting its usefulness.[206]

[49] Furthermore, while transparency reports will alert law enforcement to platforms who are failing to implement adequate moderation systems, transparency will be of limited use when detecting *actual* AVM removal failures in practice. Indeed, a platform that has stringent moderation systems in place and has been successful at removing AVM in the past, may

[201] *See* Macdonald et al., *supra* note 103, at 193–94.

[202] Bloch-Wehba, *supra* note 6, at 83.

[203] *Guarding the Guardians, supra* note 6, at 674–75.

[204] *Id.*

[205] Michael L. Rich, *Machine Learning, Automated Suspicion Algorithms, and the Fourth Amendment,* 164 U. PA. L. REV. 871, 886 (2016).

[206] *Guarding the Guardians, supra* note 6, at 674–75.

nonetheless fail to "expeditiously" remove AVM in the future if their systems are inadequate.[207] Thus, while transparency reporting eases the detection burden to some degree by providing law enforcement with additional information, it does not completely resolve the issue. As long as the AVM Act criminalizes a failure to remove AVM, rather than a failure to implement appropriate moderation systems, it will remain difficult to enforce, regardless of additional transparency requirements. Implemented alone, transparency does not address the AVM Act's underlying issues. Additional reforms are necessary to maximize overall utility in the long-term.

[50] In addition to transparency reporting, the AVM Act could also require social media companies to implement mandatory appeals mechanisms, akin to those required under the *EU Copyright Directive*.[208] This measure opens the possibility of reinstating socially-beneficial content that has been wrongfully removed. Theoretically, introducing mandatory appeals processes will provide a vital solution for the AVM Act's over-removal issue.[209]

[51] Nevertheless, this measure has a variety of drawbacks, illustrated by appeals mechanisms already offered by several social media companies. Myers-West's analysis of the current appeals mechanisms offered by Facebook, YouTube, and Twitter, reveals they are often ineffective.[210] These platforms automate their appeals processes to manage the large volume of appeals.[211] However, these algorithms are often unsatisfactory for users when examining content that falls within the "grey areas" of

[207] *See generally* Douek, *supra* note 12, at 49–52.

[208] EU Copyright Directive, *supra* note 152 at 120–21.

[209] Bridy, *supra* note 98, at 225, 227.

[210] Sarah Myers-West, *Censored, Suspended, Shadowbanned: User Interpretations of Content Moderation on Social Media Platforms*, 20 NEW MEDIA SOC. 4366, 4380 (2018).

[211] *Id.* at 4377; Bloch-Wehba, *supra* note 6, at 55.

legality.[212] Even with the opportunity for review, a large proportion of content is never reinstated.[213]

[52] This presents an even greater issue for an appeals process targeted towards AVM. Algorithms will have extreme difficulty conducting the contextual analysis necessary to determine if a piece of content constitutes AVM.[214] Accordingly, an automated appeals algorithm is unlikely to be any better at distinguishing legitimate content from AVM than the moderation algorithm that originally flagged the content would be.[215] The AVM Act could mandate that appeals should be undertaken by a human, as required under the *EU General Data Protection Regulation*.[216] However, this is unfeasible given the scale of content to be reviewed.[217] In many cases, content removal decisions made under the AVM Act will be difficult to successfully appeal.

[53] Moreover, even with mandatory appeals mechanisms, platforms who are committed to compliance remain heavily incentivized to "err on the side of caution" when deciding whether to reinstate potential AVM in order to avoid strict criminal penalties.[218] Because the AVM Act criminalizes the failure to remove AVM, there is no guarantee that mandatory appeals mechanisms will be sufficiently effective at preserving socially-beneficial

[212] Myers-West, *supra* note 210, at 4380.

[213] *See* EU Copyright Directive, *supra* note 152 at 92; Bridy, *supra* note 98, at 226–27, n. 200.

[214] *See supra* Section III (A).

[215] *See generally* Bridy, *supra* note 98, at 225–226.

[216] Council Regulation No. 2016/679, of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC, 2016 O.J. (L 119) 1, 46; *see, e.g.*, Bridy, *supra* note 98, at 225–27; Castets-Renard, *supra* note 106, at 30–31.

[217] Bloch-Wehba, *supra* note 6, at 55.

[218] Douek, *supra* note 12, at 47–48; *Free Speech is a Triangle*, *supra* note 6, at 2017.

speech. While mandatory appeals mechanisms may facilitate the reinstatement of socially-beneficial content in some cases, relying solely on appeals mechanisms is inadequate to preserve socially-beneficial speech overall.

B. Criminalizing Inadequate Systems, Not Inadequate Removal

[54] In critiquing the AVM Act and evaluating possible baseline reforms, it is clear that the AVM Act's core issues stem from the criminalization of the failure to remove AVM. Accordingly, it is suggested that the AVM Act should be amended to instead criminalize the failure to implement appropriate moderation systems.[219] This amendment will significantly increase the AVM Act's overall utility.

[55] There are two main benefits of amending the AVM Act to criminalize the failure to implement appropriate moderation systems. First, law enforcement will no longer need to undertake the difficult task of monitoring the content uploaded to platforms. Rather, they will only be required to focus on the moderation systems, which are clearly disclosed through mandatory transparency reports.[220] Transgressions become much easier to identify, increasing certainty of punishment.[221] If the penalties for failure to comply remain as they are now, the amended AVM Act will create an extremely strong incentive for rational social media platforms and hosting services to comply.[222] Even those who are less rational will be

[219] *See* Douek, *supra* note 12, at 52–53.

[220] *See* DEP'T OF COMM. AND THE ARTS, ONLINE SAFETY LEGISLATIVE REFORM 8 (2019) (Austl.) (discussing how mandatory transparency reports will provide date on harmful content, the volume of content found to breach policies, and the actions taken by the social media company).

[221] *See generally* Paternoster, *supra* note 38, at 766 (discussing how criminal sanctions are an effective deterrent to criminal behavior); Nagin and Pogarsky, *supra* note 71, at 865 (discussing the value in punishment as an effective deterrent and seeking to advance this idea with a new intellectual model).

[222] *See* Paternoster, *supra* note 38, at 817.

effectively forced to implement the prescribed moderation measures under the threat of significant and certain financial burden.[223] This offers a more satisfactory enforcement solution when compared to the current AVM Act.

[56] The second benefit is that, depending on the exact systems required, social media and hosting services will have a reduced incentive to remove lawful content.[224] Social media platforms will no longer need to weigh the risk of allowing a particular piece of questionable content to remain online against the threat of criminal sanction. As long as their moderation systems are deemed "adequate," they need not fear liability if a moderation error occurs. Moreover, because the details of the moderation systems are no longer controlled by the platforms themselves, they can be designed to foster accountability and protect legitimate speech wherever possible.[225]

[57] It would be useful to enact a regulatory body to determine what constitutes an "appropriate" moderation system. The UK *Online Harms White Paper* provides some guidance as to how this may be implemented.[226] It envisions that an independent regulator will issue codes of practice, explaining how social media companies should fulfil their duty of care to protect users from harmful content.[227] Additional guidance can be gained from the French Social Media Regulatory proposal, which recommends an administrative authority tasked with issuing compliance-based legislation to foster "accountability by design" and dialogue between key stakeholders.[228]

[223] *See Abhorrent Materials Act*, *supra* note 1, §§ 474.31, 474.34 (showing that noncompliance with the statute will result in severe financial penalties).

[224] *See* Douek, *supra* note 12, at 52.

[225] REPUBLIC OF FR., *supra* note 187, at 3.

[226] ONLINE HARMS WHITE PAPER, *supra* note 198, at 9.

[227] *Id.*

[228] REPUBLIC OF FR., *supra* note 187, at 13, 23.

[58] A similar approach could be adopted under the AVM Act, by establishing a specialist regulatory body to mandate the moderation systems social media and hosting companies must implement. This ensures the AVM Act remains technologically neutral and that it can adapt alongside developments in content moderation technology.[229] Similar to the French Government proposal, the regulatory body should oversee the protection of socially-beneficial speech by fostering accountability, transparency, and manageable appeals processes.[230] To maintain social media companies' ability to innovate, they should not be solely limited to the recommendations proposed by the regulatory body.[231] However, similar to the UK's *Online Harms White Paper*, these platforms should explain *how* their system would be more appropriate to reduce AVM and protect legitimate speech.[232]

[59] The specialist regulatory body's overall long-term aim should be to implement measures that enable more *accurate* content moderation.[233] Thus, it should conduct research into developing content moderation systems that are able to filter AVM with fewer impacts on legitimate speech. However, even in the interim, the regulatory body should attempt to implement measures that have minimum speech impacts, wherever possible. For example, instead of banning the Facebook Live feature that was used to stream the Christchurch attack, the regulator might recommend limiting its availability to only allow established, validated accounts to livestream video.[234] This measure would preserve public-interest uses of the

[229] *See generally* Roger Brownsword, *The Shaping of Our Online Worlds: Getting the Regulatory Environment Right*, 20 INT'L. J.L. & INFO. TECH. 249, 264–65 (2012) (discussing the importance of neutral drafting).

[230] *See* REPUBLIC OF FR., *supra* note 187, at 19–20, 23.

[231] *See* ONLINE HARMS WHITE PAPER, *supra* note 198, at 7.

[232] *See id.*

[233] *See, e.g.*, Bloch-Wehba, *supra* note 6, at 41(proposing auto-moderation); *see also Guarding the Guardians*, *supra* note 6, at 678.

[234] *See* Douek, *supra* note 12, at 59.

service while significantly reducing the likelihood of a perpetrator co-opting the platform to livestream AVM.[235]

[60] Nevertheless, some limitations to this proposal remain. Before more accurate systems can be developed, there will always be a difficult struggle between ensuring the removal of AVM and maintaining legitimate speech. As Douek notes, errors are inevitable.[236] Regardless, transparency and accountability will enable effective dialogue between users, the regulatory body, and social media companies.[237] When errors occur, stakeholders can discuss and research solutions to produce improved systems.

[61] Compared to the present approach, this recommendation offers a more measured solution that aligns with the UPC by producing a net social benefit. It provides the tools necessary to facilitate the removal of AVM, while minimizing collateral removal of lawful speech. Accordingly, it is necessary to amend the AVM Act to regulate AVM on social media more appropriately.

VI. CONCLUSION

[62] The AVM Act should be amended to regulate AVM more appropriately on social media. Applying Deterrence Theory suggests that the current AVM Act is unconducive to AVM removal due to enforcement inefficiencies that lower certainty of punishment.[238] Additionally, the AVM Act unnecessarily induces the over-removal of legitimate socially-beneficial speech online, which significantly reduces its overall utility[239]. Thus, the AVM Act fails to align with the utilitarian principle of criminalization. To ameliorate this concern, the AVM Act should be amended to include mandatory transparency reporting and appeals

[235] *Id.*

[236] *Id.* at 48.

[237] *See* REPUBLIC OF FR., *supra* note 187, at 22.

[238] *See supra* Section III.

[239] *See id.*

mechanisms. However, without further measures, these reforms will only produce minimal additional utility. Therefore, the AVM Act should be further amended to criminalize the failure to implement appropriate content moderation systems, rather than the failure to remove AVM. An independent regulatory body should be established to determine what an appropriate content moderation system should require, in light of the importance of socially-beneficial speech. Ultimately, this approach is more conducive to facilitating a reduction in AVM, while minimizing collateral over-removal of legitimate speech. The result is a law that increases social utility, and effectively upholds the UPC.