A PORTFOLIO APPROACH TO TRUSTED INTERMEDIARIES FOR ONLINE CONTENT AND CONDUCT

DAVID SULLIVAN*

Proponents of U.S. online safety legislation frequently argue that the costs of inaction outweigh whatever negative consequences might arise from altering the status quo. This position does not account for the full range of industry and multistakeholder initiatives that have developed, enabled by the First Amendment and Section 230, to contend with harmful content. This article assesses case studies of these entrepreneurial efforts based on criteria for trusted intermediaries: effectiveness, legitimacy, and accountability. It shows how a portfolio approach of partial solutions can evolve into a holistic approach to online trust and safety, one that will be of critical importance as courts continue to deliberate the extent to which governments may regulate company decision-making regarding online content and conduct.

[*] David Sullivan is the executive director of the Digital Trust and Safety Partnership, writing in his personal capacity. This article revises and expands upon remarks given during the "Regulation of Lies, Misinformation, and Deepfakes" panel at the Silicon Flatirons Flagship Conference in February 2024, and responds to comments made during other panels, particularly the interchange between Phil Weiser and Susan Ness during the "Global Fractures in Technology Policy" panel, where they discussed the need for trusted intermediaries on challenging cross-border issues such as AI and content moderation. My thanks to Farzaneh Badiei, Daphne Keller, Jess Miers, Blake Reid, Matthew Schruers, and Ali Sternburg for illuminating conversations and comments that informed this article.

INTRODUCTION

Concerns about harmful online content and conduct have been top of mind for all branches of the U.S. federal government,[1] dozens of U.S. states,[2] and governments around the world.[3] Despite the increased appetite for internet regulation, there has been relatively little movement in Congress, notwithstanding rapid changes at the state level and internationally. Bipartisan concern belies deep disagreement over what should be done. Republicans generally favor laws that restrict the ability of companies to moderate content on their services. Democrats tend to promote regulations that would incentivize companies to remove more content, particularly hate speech, disinformation, and misinformation.[4] Constitutional

[1] Illustratively, dozens of bills seeking to reform Section 230 have been introduced in recent years as tracked by Lawfare. *Section 230 Legislation Tracker*, LAWFARE (Sep. 19, 2023) https://www.lawfaremedia.org/projects-series/section-230-tracker [https://perma.cc/VX9R-CESS]. The Biden-Harris Administration established a White House Task Force to Address Online Harassment and Abuse which published its final report in 2024, as well as a Task Force on Kids Online Health & Safety. *See* THE WHITE HOUSE, WHITE HOUSE TASK FORCE TO ADDRESS ONLINE HARASSMENT AND ABUSE (2024); *see also*, *Kids Online Health and Safety*, NAT'L TELECOMM. AND INFO. ADMIN., https://www.ntia.gov/programs-and-initiatives/kids-online-health-and-safety [https://perma.cc/YLF2-9W29] (last visited Nov. 1, 2024). Finally, the Supreme Court has focused on issues related to harmful content and behavior in recent years. *See Gonzalez v. Google LLC*, 598 U.S. 617 (2023); *Twitter, Inc. v. Taamneh*, 598 U.S. 471 (2023); *Moody v. NetChoice, LLC*, 144 S.Ct. 2383 (2024); *Murthy v. Missouri*, 144 S.Ct. 1972 (2024).

[2] According to the National Conference of State Legislatures, 35 states addressed legislation targeting social media and children in 2023, with 13 states enacting laws or passing resolutions. *Social Media and Children 2023 Legislation*, NAT'L CONF. OF STATE LEGISLATURES, https://www.ncsl.org/technology-and-communication/social-media-and-children-2023-legislation [https://perma.cc/EU2B-355Z] (last updated Jan. 26, 2024).

[3] Internationally, significant content regulations enacted in recent years include Australia's Online Safety Act 2021, Singapore's Online Safety Act, the European Union's Digital Services Act, the United Kingdom's Online Safety Act 2023, and India's IT Amendment Rules 2022, among others. Dozens of countries are actively considering legislation related to content in addition to these examples. *See Regional Activity, Policy changes between 31 Dec 2019 and 11 Jun 2024*, DIGITAL POLICY ALERT, https://digitalpolicyalert.org/policy-area/content-moderation?period=2020-01-01,2024-06-12#regional-activity [https://perma.cc/NZR7-7B2T] (last visited Sep. 7, 2024).

[4] Danielle Citron & Quinta Jurecic, *FOSTA's Mess*, VA J. OF L. & TECH., Spring 2023, at 1–15. There has been extensive analysis of the general Congressional dynamics on tech policy. *See* Brian Fung, *Congress hasn't been able to make social media safer.*

hurdles on speech regulation pose obstacles to even the most modest regulatory proposals for both systems and process and transparency regulations.[5]

This article proposes that a wide range of industry and multistakeholder collaborative initiatives are already providing *de facto* solutions to online content concerns despite the absence or unconstitutionality of *de jure* regulations. Skeptics often dismiss such initiatives as public relations or a means of deterring legislation or regulation. Such cynicism tends to overlook the necessity of industry-wide solutions to address industry-wide problems, especially in the case of cross-platform abuse, where bad actors use a combination of services to inflict harm.

Given public skepticism toward technology companies, it is important to rigorously distinguish credible and legitimate initiatives from those that are not. Using the criteria for trusted intermediaries identified by Philip J. Weiser, this article identifies the strengths and weaknesses of a range of entrepreneurial efforts across the product development lifecycle.[6] Through these case studies, this article argues that policymakers and practitioners should adopt a "portfolio approach"[7] to problematic online content

Here's why, CNN BUSINESS (Feb. 1, 2024, 6:01 PM), https://www.cnn.com/2024/02/01/tech/social-media-regulation-bipartisan-support/index.html [https://perma.cc/9CRC-MYFV]. There are examples of bipartisan legislative initiatives, including the Kids Online Safety and Privacy Act (KOSPA), which passed the Senate with only three Senators voting against it, but faces obstacles in the House of Representatives. *See* Lauren Feiner, *Senate passes the Kids Online Safety Act*, THE VERGE (Jul. 30, 2024, 11:05 AM), https://www.theverge.com/2024/7/30/24205718/senate-passes-kids-online-safety-act-kosa-content-moderation [https://perma.cc/ZTX2-BHA6]. However, even where there is bipartisan support for legislation, Republicans and Democrats differ significantly on the goals and implementation of such legislation. *See* Kris Kobach, *KOBACH: Liberals Hijack Online Child Safety Bill, Handing Khan-trol to FTC*, DAILY CALLER (May 20, 2024, 11:31 AM), https://dailycaller.com/2024/05/20/kobach-kosa-lgbtq-kids-online-safety-act-parental-surveillance [https://perma.cc/YL7R-LD4Y].

[5] Eric Goldman, *The Constitutionality of Mandating Editorial Transparency*, 73 HASTINGS L. J. 1203 (2022).

[6] Weiser examines public and private regulatory experimentation and identifies "three principal criteria for regulatory innovation," which are 1) effectiveness; 2) legitimacy and adherence to public norms; and 3) accountability. Philip J. Weiser, *Entrepreneurial Administration*, 97 B.U. L. REVIEW 2011, 2037 (2017).

[7] In an empirical study of policy responses to disinformation, Bateman and Jackson assessed 10 interventions to counter disinformation and concluded that "none… were simultaneously well-studied, very effective, and easy to scale." Instead, they recommended that "[p]olicymakers should act like investors, pursuing a diversified mixture of counter-disinformation efforts while learning and rebalancing over time." Jon Bateman & Dean Jackson, *Countering Disinformation Effectively: An Evidence-Based Policy Guide*, CARNEGIE ENDOWMENT FOR INT'L PEACE 2 (Jan. 31, 2024), https://carnegie-production-assets.s3.amazonaws.com/static/files/Carnegie_Countering_Disinformation_Effectively.pdf [https://perma.cc/2HTC-8ZPF].

and behavior, including disinformation, deepfakes,[8] and technology-facilitated gender-based violence. By taking rigorous stock of the diverse assets already in our problematic content policy portfolio,[9] we can understand whether current approaches are balanced, risk-appropriate, and suited to deliver the returns that society seeks. This will provide an evidence base that will improve the quality of public debate about whether legal reforms might be warranted, and if so, what they should encompass.

I. CHALLENGES WITH CONTENT REGULATION

Digital products and services facilitate the full spectrum of human activity, including enabling expression, communication, and access to information. Although broadly beneficial, such services also enable users to create and share content and engage in behavior that is unwanted, abusive, or illegal.

In the United States, the inherently expressive nature of such services creates a high bar for regulation, given First Amendment restrictions. Although dozens of bills have been introduced to Congress in recent years proposing varying approaches to regulation,[10] none have been passed as of the publication of this article since the Fighting Online Sex Trafficking Act of 2018.[11] This has not stopped a flurry of legislation enacted at the state level, much of which faces litigation stemming in large part from First Amendment concerns.[12]

Internationally, regulation has moved faster and with more consequence due to the passage of the Digital Services Act in the

[8] According to the Digital Trust & Safety Partnerships glossary of trust and safety terminology, "[a] deepfake is a form of synthetic media where an image or recording is altered to misrepresent someone doing or saying something that was not done or said." *Digital Trust & Safety Glossary of Terms*, DIGITAL TRUST & SAFETY PARTNERSHIP 22 (2023), https://dtspartnership.org/wp-content/uploads/2023/07/DTSP_Trust-Safety-Glossary_July-2023.pdf [https://perma.cc/8VBX-V76V].

[9] This article is concerned with collective efforts to address harmful content across industry and through multistakeholder initiatives. For a comprehensive taxonomy of options that individual companies use to remedy content or conduct violations, *see* Eric Goldman, *Content Moderation Remedies*, 28 MICH. TECH. L. REV. 1 (2021).

[10] *See generally Section 230 Legislation Tracker*, *supra* note 1.

[11] *See* Citron & Jurecic, *supra* note 4.

[12] In addition to *Moody v. NetChoice, LLC*, 144 S.Ct. 2383 (2024), preliminary injunctions have been issued in response to litigation in California, Arkansas, Ohio, and New York. *See NetChoice, LLC v. Bonta*, 692 F. Supp. 3d 924 (N.D. Cal. 2023), *aff'd in part, vacated in part,* No. 23-2969, 2024 WL 3838423 (9th Cir. Aug. 16, 2024); *NetChoice, LLC v. Griffin*, No. 5:23-CV-05105, 2023 WL 5660155 (W.D. Ark. Aug. 31, 2023); *NetChoice, LLC v. Yost*, No. 2:24-CV-00047, 2024 WL 555904 (S.D. Ohio Feb. 12, 2024); *Volokh v. James*, 656 F. Supp. 3d 431 (S.D.N.Y. 2023).

European Union.[13] Additionally, various forms of online safety laws in Australia, the United Kingdom, and Singapore are among dozens of laws that have been proposed or enacted worldwide.[14]

The result is a fragmented and incoherent global regulatory regime for digital products and services. Companies are complying with laws in international jurisdictions that, if implemented in the United States, on the spectrum of First Amendment concerns, would range from flagrantly unconstitutional to "it's complicated." For example, data protection impact assessments (DPIAs) are a key component of the EU General Data Protection Regulation (GDPR)[15] and are also required under the UK Age Appropriate Design Code.[16] A modified DPIA requirement was included in California's Age Appropriate Design Code,[17] but was struck down by the 9th Circuit who said it "clearly compels speech by requiring covered businesses to opine on potential harms to children."[18] In other cases, international regulatory requirements are creating additional work for companies' internal compliance teams, but not necessarily leading to better safety outcomes. For example, transparency reporting was a voluntary effort by companies that is now mandated under multiple online safety regimes. As a result, companies are having to produce bespoke transparency reports for multiple jurisdictions, without a clear effect on safety.[19]

Moreover, legislative and regulatory proposals pertaining to digital services are intended to deal with many types of problematic content and behavior, including content deemed illegal in one or

[13] *See generally* Commission Regulation 2022/2065 of Oct. 19, 2022, Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act), 2022 O.J. (L 277) 1.

[14] *See Regional Activity, Policy changes between 31 Dec 2019 and 11 Jun 2024, supra* note 3.

[15] *See generally* Commission Regulation 2016/679 of Apr. 27 2016, 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 (General Data Protection Regulation), 2016 O.J. (L 119) 1 [hereinafter GDPR] (Article 35 covers DPIAs).

[16] Elizabeth Denham, *Age appropriate design: a code of practice for online services*, INFO. COMM'NS OFF., https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/childrens-information/childrens-code-guidance-and-resources/age-appropriate-design-a-code-of-practice-for-online-services/2-data-protection-impact-assessments [https://perma.cc/V8MH-TWYV] (last visited Nov. 17, 2024).

[17] Assemb. B. No. 2273, 2022 Cal. Gen. Assemb., (Cal. 2022).

[18] *NetChoice, LLC v. Bonta*, No. 23-2969, 2024 WL 3838423, at *9 (9th Cir. Aug. 16, 2024).

[19] Illustratively, Meta now has a page for Regulatory and Other Transparency Reports that includes 17 different reports. Including ten EU reports, as well as reporting for Austria, Brazil, Korea, India, Germany, Türkiye, and Norway. *See generally Regulatory and Other Transparency Reports*, META, https://transparency.meta.com/reports/regulatory-transparency-reports [https://perma.cc/5TE6-WNHQ] (last visited Dec. 23, 2024).

more jurisdictions, as well as broader "lawful but awful"[20] content. These regulations often encourage a proportionate and risk-based approach, stating that companies should balance safety considerations against human rights, such as freedom of expression; however, in practice, their effect is likely to be more blunt. Ultimately, companies will be incentivized to err on the side of overcompliance, removing more content than is legally required.[21]

II. *DE FACTO* REGULATION AND TRUSTED INTERMEDIARIES

Politicians around the world have frequently called the Internet a "digital wild west,"[22] arguing that digital services have benefited from "self-policing and self-regulation,"[23] and proclaiming that "the era of self-regulation is over."[24] But claims that internet services are somehow lawless and unregulated are routinely overstated.

As Eric Goldman lays out, in the United States, regulation of internet services is governed by the First Amendment and Section 230.[25] Aside from narrowly defined areas not protected by the First Amendment, such as child sexual abuse imagery and incitements to violence, much of the content and behavior that politicians decry on digital services is constitutionally protected speech. The liability protection created by Section 230 has statutory exemptions for

[20] Daphne Keller, *Lawful but Awful? Control over Legal Speech by Platforms, Governments, and Internet Users*, UNIV. OF CHI. L. REV. ONLINE ARCHIVE (June 28, 2022), https://lawreviewblog.uchicago.edu/2022/06/28/keller-control-over-speech [https://perma.cc/M4LL-HDY9].

[21] *Preventing "Torrents of Hate" or Stifling Free Expression Online: An Assessment of Social Media Content Removal in France, Germany, and Sweden*, THE FUTURE OF FREE SPEECH (May 2024), https://futurefreespeech.org/wp-content/uploads/2024/05/Preventing-Torrents-of-Hate-or-Stifling-Free-Expression-Online-The-Future-of-Free-Speech.pdf [https://perma.cc/LW7L-EJ7L].

[22] Thierry Breton (@ThierryBreton), X, (Jan. 19, 2022, 6 :00AM), https://x.com/ThierryBreton/status/1483786510214303744?mx=2 [https://perma.cc/Z8PP-LD9T].

[23] Richard Blumenthal, *Blumenthal Chairs Hearing With Head of Instagram on Social Media's Dangers to Kids & Legislative Solutions*, RICHARD BLUMENTHAL U.S. SENATOR FOR CONN. (Dec. 9, 2021), https://www.blumenthal.senate.gov/newsroom/press/release/blumenthal-chairs-hearing [https://perma.cc/9S9A-LF7V].

[24] Jan Schakowsky, *Schakowsky Statement on Facebook Whistleblower Testimony*, UNITED STATES CONGRESSWOMAN JAN SCHAWKOWKY (Oct. 7, 2021), https://schakowsky.house.gov/media/press-releases/schakowsky-statement-facebook-whistleblower-testimony [https://perma.cc/8Q8Y-4KKZ].

[25] Eric Goldman, *The United States' Approach to 'Platform' Regulation* 4 (Santa Clara U. Legal Studies Research, Paper No. 4404374, 2023), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4404374 [https://perma.cc/MQP5-YWHB].

federal criminal law (which includes things like child endangerment, narcotics trafficking, and internet gambling), as well as intellectual property and state criminal prosecutions related to sex trafficking. Moreover, recent jurisprudence suggests a common law trend,[26] where Section 230 protections do not apply when the harm emanates from the design of the product.[27]

Another purpose of Section 230 was to "encourage service providers to self-regulate the dissemination of offensive material over their services."[28] The type of self-regulation associated with Section 230 is typically considered at an individual company level by providing liability protection for their content moderation decisions. However, the liability protections provided by Section 230 have also contributed to the development of broader industry and multistakeholder efforts to address harmful content. Within the boundaries created by the First Amendment and the federal and state laws described above, a complex system of soft law, industry best practices, and multistakeholder partnerships, referred to as "networked governance," has emerged.[29] Kate Klonick, while questioning the sufficiency of these efforts and suggesting some of them should be mandatory, has referred to this constellation of largely voluntary efforts as "the Golden Age of Tech Accountability."[30]

Notably, many instances of *de facto* regulation employ some version of the trusted intermediary concept, achieved via criteria described by Weiser as applied to private entities: effectiveness ("whether it advances its envisioned purposes effectively"), legitimacy, adherence to public norms ("the best practice is to ensure they operate openly and transparently"), and accountability

[26] *See* William Stevens, *The Common Law Origins of the Infield Fly Rule*, 123 U. OF PA. L. REV. 1474 (1975).

[27] *Lemmon v. Snap, Inc.*, 995 F.3d 1085 (9th Cir. 2021).

[28] *Zeran v. Am. Online, Inc.*, 129 F.3d 327, 331 (4th Cir. 1997).

[29] Robyn Caplan, *Networked Governance*, THE YALE-WIKIMEDIA INITIATIVE ON INTERMEDIARIES & INFO., Aug. 2022, at 6. ("Networked governance is useful as a framework for studying platform governance, particularly in tracing how platform companies make use of external stakeholders, such as civil society organizations and academics, in the development of platform policies, such as in the setting of community guidelines."). The term "platform" is also contentious. As Robert Gorwa writes, while defending his use of the term, "Overall, 'platform' is an imperfect and ambiguous term, one that rolls off the tongue of some while confusing others." This article prefers the term "digital product or service" which applies to a broader set of services and avoids some of the ambiguities associated with platform, which it uses only when referencing others' work. *See* ROBERT GORWA, THE POLITICS OF PLATFORM REGULATION: HOW GOVERNMENTS SHAPE ONLINE CONTENT MODERATION (Oxford University Press, 2024).

[30] Kate Klonick, *The End of the Golden Age of Tech Accountability*, THE KLONICKLES (Mar. 3, 2023), https://klonick.substack.com/p/the-end-of-the-golden-age-of-tech [https://perma.cc/T9BU-QJ4F].

("auditing, certification, and oversight regimes to encourage compliance").[31]

III. A PORTFOLIO OF ENTREPRENEURIAL INITIATIVES ACROSS THE PRODUCT DEVELOPMENT LIFECYCLE

Adapting the portfolio approach,[32] this article explores a broader set of industry and multistakeholder collaborative efforts that are attempting to tackle different facets of the complex challenge presented by misinformation, deepfakes, and other abusive online content.

Each example pertains to one of the five commitments in the Best Practices Framework set out by the Digital Trust & Safety Partnership (DTSP), [33] an organization led by the author. DTSP is a voluntary industry partnership that aims to bring together technology companies that provide diverse digital products and services around a shared framework of best practices for trust and safety.[34] The DTSP commitments mirror the product development lifecycle and provide a means of organizing discussion about trust and safety across five distinct areas: product development, governance, enforcement, improvement, and transparency. The case studies presented here are not explicitly referred to within the DTSP Best Practices Framework but provide illustrative examples of how companies might implement their commitments with regard to misinformation and deepfake risks.

This next part explores these case studies, using the "three criteria for sound institutional design and regulatory experimentation" identified by Weiser,[35] by examining proven effectiveness, legitimacy, and accountability.

[31] Weiser, *supra* note 6, at 2037, 2045.

[32] Bateman, *supra* note 7.

[33] *Digital Trust & Safety Partnership Best Practices Framework*, DIGITAL TRUST & SAFETY PARTNERSHIP, https://dtspartnership.org/best-practices [https://perma.cc/B7QX-LMKC] (last visited Nov. 1, 2024).

[34] DTSP does not determine what types of content or conduct are suitable for the products and services offered by partner companies but aims to describe the practices used by partners as part of their trust and safety operations, and to identify best practices that can be rigorously assessed as to their effectiveness. As the author has written, "Whereas other efforts to address trust and safety practices often start from theoretical approaches that are then applied operationally, our approach has been to begin by describing how practitioners understand the terms they use and how this informs their practices." Farzaneh Badiei et al., *Toward a Common Baseline Understanding of Trust and Safety Terminology*, J. OF ONLINE TR. AND SAFETY, 2023, at 1.

[35] Weiser *supra* note 6.

A. *Technical standards for provenance as part of safety-by-design*

The concept of deliberately designing digital products to account for risks has spread from privacy to security to safety.[36] Although safety-by-design includes elements that span the full product lifecycle, it centers around product development and ensuring safety is not an afterthought in that process.[37]

Can safety-by-design approaches mitigate the risks to democracy posed by deepfakes? Elections that occurred in 2024 in the United States and around the world have drawn attention to the potential for deepfakes to influence democratic practices, especially elections.[38] As a result, industry efforts to establish and track the provenance of digital data have taken on new urgency.[39] Once obscure technical standards development processes have

[36] Information security is "[t]he protection of information and information systems from unauthorized access, use, disclosure, disruption, modification, or destruction in order to provide confidentiality, integrity, and availability." Glossary Definition of *INFOSEC*, NIST COMPUT. SEC. RES. CTR., https://csrc.nist.gov/glossary/term/infosec [https://perma.cc/35GJ-KLTG] (last visited Nov. 1, 2024). Trust and safety is "[t]he field and practices employed by digital services to manage content- and conduct-related risks to users and others, mitigate online or other forms of technology-facilitated abuse, advocate for user rights, and protect brand safety." *Trust & Safety Glossary of Terms*, DIGIT. TR. & SAFETY P'SHIP 12 (2023), https://dtspartnership.org/wp-content/uploads/2023/07/DTSP_Trust-Safety-Glossary_July-2023.pdf [https://perma.cc/8VBX-V76V]. In the first issue of the Journal of Online Trust and Safety, editors Elena Cryst, Shelby Grossman, Jeff Hancock, Alex Stamos, and David Thiel write: "[t]rust and safety is the study of how people abuse the internet to cause real human harm, often using products the way they are designed to work. If someone uses a peer-to-peer messaging app to send a message that threatens to hurt the recipient's family, the product is being used as intended, namely, to send a message. However, the message content itself is causing harm." Elena Cryst et al., *Introducing the Journal of Online Trust and Safety*, 1 J. ONLINE TR. & SAFETY 1, 1 (2021). For more on how "safety by design" builds on the concepts of privacy by design and security by design, *see* John Perrino, *Using 'Safety by Design' to Address Online Harms*, BROOKINGS (July 26, 2022), https://www.brookings.edu/articles/using-safety-by-design-to-address-online-harms [https://perma.cc/D7EJ-BNAL].

[37] *Safety by Design*, TRUST & SAFETY PROFESSIONAL ASSOCIATION, https://www.tspa.org/curriculum/ts-curriculum/safety-by-design [https://perma.cc/SP3E-SN75] (last visited Nov. 1, 2024).

[38] Victoria Elliott & Makena Kelly, *The Biden Deepfake Robocall Is Only the Beginning*, WIRED (Jan. 23, 2024, 12:58 PM), https://www.wired.com/story/biden-robocall-deepfake-danger [https://perma.cc/8MAR-L3TF].

[39] Elections are not the only reason for the urgent focus on this issue. The proliferation of online tools that allow the trivial creation of synthetic media, including deepfake images used for harassment and abuse, have put this issue in the spotlight. New research from the nonprofit Thorn states 1 in 10 minors report that they know of cases where their friends and classmates have created synthetic non-consensual intimate images (or "deepfake nudes") of other kids using generative AI tools. THORN, YOUTH PERSPECTIVES ON ONLINE SAFETY, 2023 (Aug. 2024), https://info.thorn.org/hubfs/Research/Thorn_23_YouthMonitoring_Report.pdf [https://perma.cc/6V9R-7VPE].

become routinely incorporated into major announcements by technology companies.[40]

There are two overarching approaches to distinguishing authentic, human-created content, from synthetic, AI-generated content. The first is technology that can detect whether content was created by generative AI. Such tools face substantial challenges with accuracy and reliability.[41] The second is the opt-in approach, where content creators provide data about how the content was created and how it has changed.[42] This case concerns the latter approach, which is referred to as content provenance.

In recent years, several organizations have worked together to develop content provenance standards. The Coalition for Content Provenance and Authenticity (C2PA) and the related Content Authenticity Initiative (CAI) provide an example of how industry collaboration in the formation of technical standards can be used to further an overall commitment to safety-by-design. C2PA is the formal standards development organization creating technical standards to embed data about the source and history of media content, which they term "content credentials."[43] C2PA is led by Adobe, Arm, Intel, Microsoft, and Truepic.[44] CAI is a broader cross-industry community exploring implementation of the standards and the development of open source tools.[45]

Are these provenance standards effective? Similar to the Internet Engineering Task Force's cardinal principle of "rough consensus and running code,"[46] the effectiveness of C2PA standards can be measured based on real world implementation of

[40] *See, e.g.,* BRAD SMITH, PROTECTING THE PUBLIC FROM ABUSIVE AI-GENERATED CONTENT, MICROSOFT (2024), https://cdn-dynmedia-1.microsoft.com/is/content/microsoftcorp/microsoft/msc/documents/presentations/CSR/Protecting-Public-Abusive-AI-Generated-Content.pdf [https://perma.cc/P5M6-HE94].

[41] Stuart A. Thompson & Tiffany Hsu, *How Easy Is It to Fool A.I.-Detection Tools?*, N.Y. TIMES (June 28, 2023), https://www.nytimes.com/interactive/2023/06/28/technology/ai-detection-midjourney-stable-diffusion-dalle.html [https://perma.cc/96MP-UF6M].

[42] Tate Ryan-Mosley, *The Inside Scoop on Watermarking and Content Authentication*, MIT TECH. REV., (Nov. 6, 2023), https://www.technologyreview.com/2023/11/06/1082996/the-inside-scoop-on-watermarking-and-content-authentication [https://perma.cc/47TC-JFV9].

[43] CONTENT CREDENTIALS, https://contentcredentials.org [https://perma.cc/JF3R-T7GV] (last visited Nov. 1, 2024).

[44] *C2PA Membership*, COALITION FOR CONTENT PROVENANCE AND AUTHENTICITY, https://c2pa.org/membership [https://perma.cc/QH3B-U9JL] (last visited Nov. 1, 2024).

[45] *How it works*, CONTENT AUTHENTICITY INITIATIVE, https://contentauthenticity.org/how-it-works [https://perma.cc/CEW3-XV8R] (last visited Nov. 1, 2024).

[46] *Introduction to the IETF*, INTERNET ENGINEERING TASK FORCE, https://www.ietf.org/about/introduction [https://perma.cc/9ZHG-S6EG] (last visited Nov. 1, 2024).

their specifications. Here, there are notable successes, with content credentials being implemented in various ways by the BBC, TikTok, and even on the Leica M11-P digital camera. Meanwhile, many of the largest AI players have joined the C2PA in recent months, including Google, Meta,[47] and OpenAI. One potential effectiveness concern, which has been acknowledged by C2PA's leadership, is the gap between what content provenance can achieve versus the broader fight against misinformation writ large, where the attribution-based approach of C2PA/CAI will need to be matched with efforts to detect manipulated media and educate the public through media literacy. A second concern is that security researchers have documented ways that credentials can be stripped from content or forged "to create false attribution, impersonations, and propaganda."[48] So, while embedding data about the media's source provides a good start, if users can't tell if the content has been altered, the end goal of the technical standard still will not be met.

With incidents of election-related deepfakes on the rise, governments and companies alike have embraced the concept of watermarking AI images to prevent deception. Watermarking AI-generated images, however, is not the same as establishing provenance for authentic images. Thus, content credentials are going to have a limited impact addressing the issues of greatest short-term concern (e.g., deepfake electoral disinformation). As Jacob Hoffman-Andrews of the Electronic Frontier Foundation put it: "it's still a fiendishly complicated scheme, since the chain of verifiability has to be preserved through all software used to edit photos. And most cameras will never produce this metadata, meaning that its absence can't be used to prove a photograph is fake."[49] Ultimately, the considerable technical challenges to adopting content provenance at scale are modest compared to the size of the social challenges they ultimately seek to address. As Adobe's Andy Parsons told *Quartz:* "None of these countermeasures is a silver bullet. It's going to take government, civil society,

[47] Nick Clegg, *Labeling AI-Generated Images on Facebook, Instagram and Threads*, META (Feb. 6, 2024), https://about.fb.com/news/2024/02/labeling-ai-generated-images-on-facebook-instagram-and-threads [https://perma.cc/9WA4-A2B4].

[48] *C2PA and Untrusted Certificates*, THE HACKER FACTOR BLOG (July 8, 2024), https://www.hackerfactor.com/blog/index.php?/archives/1037-C2PA-and-Untrusted-Certificates.html [https://perma.cc/DF5E-Q9FT].

[49] Jacob Hoffman-Andrews, *AI Watermarking Won't Curb Disinformation*, ELEC. FRONTIER FOUND. (Jan. 5, 2024), https://www.eff.org/deeplinks/2024/01/ai-watermarking-wont-curb-disinformation [https://perma.cc/N4DY-3XKD].

technology companies, and a variety of technological approaches to really address misinformation."[50]

In terms of legitimacy, the broad and growing membership of C2PA and CAI is an important sign of progress. Moreover, C2PA has aligned with some of the traditional norms and standards of the development community. In particular, C2PA is a project of the Joint Development Foundation, a project of the Linux Foundation that provides a common set of industry standard membership structures, legal agreements, and other administrative requirements for standards bodies.[51] For example, C2PA's membership agreement is publicly available on their website.[52] Moreover, the participation of respected civil society organizations and experts, particularly Sam Gregory of Witness, has helped establish credibility with other stakeholders.[53]

Content credentials are still a relatively new development, so accountability structures remain relatively nascent. The ability for anyone to check content credentials provides an opportunity to guard against cheating, but with the deployment of this standard in an early stage, the amount of verifiable content remains relatively low.[54] Moreover, as long as vulnerabilities in the specification allow clearly fake images to be tampered with, there remains a real possibility of public loss of trust in C2PA. This risk has not stopped lawmakers from advancing proposals that would rely on the standard, even while it is still very much in development.[55] Although content provenance presents perhaps the

[50] Laura Bratton, *Adobe is fighting AI election deepfakes. Here's how*, QUARTZ (Apr. 18, 2024), https://qz.com/deepfake-elections-ai-misinformation-adobe-content-cred-1851417898 [https://perma.cc/TC7X-XDXR].

[51] *FAQ*, JOINT DEVELOPMENT FOUNDATION, https://jointdevelopment.org/faq [https://perma.cc/3745-VSKR] (last visited Nov. 6, 2024).

[52] *Membership Agreement*, C2PA, https://cdn.platform.linuxfoundation.org/agreements/c2pa-fund.pdf [https://perma.cc/98CA-N6F3] (last visited Nov. 6, 2024).

[53] *The Need for Transparency in Artificial Intelligence: Hearing Before the Subcomm. on Consumer Prot., Prod. Safety and Data Sec. of the S. Comm. on Com., Sci. and Transp.*, 118th Cong. 1 (2023) (statement of Sam Gregory, Executive Director, WITNESS).

[54] For example, the BBC announced use of content credentials in March 2024. *See New technology to show why images and video are genuine launches on BBC News*, BBC (Mar. 4, 2024), https://www.bbc.com/mediacentre/2024/content-credentials-bbc-verify [https://perma.cc/6R7J-9J7M]. However, as of June 12, 2024, none of the articles featured on the BBC Verify website included content credentials.

[55] For example, *see* Assemb. B. 3211, 2023-2024 Reg. Sess. (Cal. 2024), which would mandate the use of watermarking and content provenance standards by Generative AI providers. For criticism of this approach, *see* Dean w. Ball, *California should rethink its broad and sloppily drafted deepfake bill*, UNDERSTANDING AI (July 29, 2024), https://www.understandingai.org/p/california-should-rethink-its-broad [https://perma.cc/4J25-6TYH].

most promising technical contribution to the fight against AI-generated abuse, and while it is expected that any technical standard will require significant iteration to overcome flaws,[56] there is more work to be done before content provenance can achieve what policymakers, tech executives, and perhaps the wider public will expect from it. That said, the standard progresses, with the latest version (2.1) strengthened against tampering attacks,[57] and with the establishment of a Conformance Task Force within the C2PA that is actively testing conformity and certification with the standard.[58]

B. Oversight Board policy recommendations

The establishment of the Oversight Board by Meta remains one of the most ambitious voluntary efforts in the field of platform governance.[59] A detailed analysis of all aspects of the Oversight Board is beyond the scope of this article,[60] which will instead focus on one aspect of the Board's work: non-binding recommendations. These non-binding recommendations provide a mechanism for recommending changes to policy, operations, and product design that could enable more systematic changes to Meta products beyond up-or-down decisions on what content is acceptable. Public attention tends to focus on the ability of the Oversight Board to issue binding decisions in response to appeals from users of Facebook and Instagram on whether content taken down or left up was done so consistently with Meta's community guidelines and other relevant policies. But the Oversight Board also issues non-binding policy recommendations, either in conjunction with its judgements or in response to requests for policy advice from Meta. The Board reports that it has made 288 recommendations to Meta

[56] For examples of constructively critical recommendations to improve C2PA, *see* Dean W. Ball, *Deepfakes and the Art of the Possible*, HYPERDIMENSIONAL (May 30, 2024), https://www.hyperdimensional.co/p/deepfakes-and-the-art-of-the-possible [https://perma.cc/7GR9-ZFYZ].

[57] Laurie Richardson, *How we're increasing transparency for gen AI content with the C2PA*, GOOGLE (Sept. 17, 2024), https://blog.google/technology/ai/google-gen-ai-content-transparency-c2pa [https://perma.cc/L8V3-9QT3].

[58] Truepic, Comment Letter on Draft Report NIST AI 100-4, Reducing Risks Posed by Synthetic Content: An Overview of Technical Approaches to Digital Content Transparency, https://www.regulations.gov/comment/NIST-2024-0001-0032 [https://perma.cc/XV5R-N8U4].

[59] Kate Klonick, *The Facebook Oversight Board: Creating an Independent Institution to Adjudicate Online Free Expression*, 129 YALE L.J. 2418, 2418 (2020).

[60] Evelyn Douek, *The Meta Oversight Board and the Empty Promise of Legitimacy*, 37 HARV. J.L. & TECH. 373 (discussing an extensive evaluation of the Board's impact).

as of November 2024,[61] and while Meta is not obligated to accept these recommendations, the company has committed to respond publicly to them within 60 days.[62]

For example, in February 2024, the Board upheld Meta's decision to leave up a video that was manipulated to make it appear that U.S. President Biden was inappropriately touching the chest of his adult granddaughter. In conjunction with this decision, the Board recommended that Meta revise its Manipulated Media policy, "finding it to be incoherent, lacking in persuasive justification, and inappropriately focused on how content has been created, rather than on which specific harms it aims to prevent (for example, to electoral processes)."[63] In April 2024, Meta announced changes to this policy in response to the Board's recommendations.[64] This example demonstrates the potential for the Oversight Board to serve as a trusted intermediary that can inform the development and update of content policies and other product governance.

Regarding effectiveness, the Board tracks progress on these recommendations on its website and has reflected on this process through their own scholarship in a commentary published in the Journal of Online Trust and Safety.[65] The Board has also developed and published its methodology for evaluating Meta's response to its recommendations. The Board has noted that recommendations relating to content policy or transparency are more frequently documented as fully or partially implemented, "because, in most cases, their implementation requires a public-facing change (e.g., a change to the public Community Standards or to user notifications) that directly adopts the language of the recommendation."[66] In contrast, recommendations regarding enforcement and other non-public facing aspects are more challenging to independently verify.

[61] *See Latest Implementation Assessment*, OVERSIGHT BOARD, https://www.oversightboard.com/explore-our-recommendation-tracker [https://perma.cc/F5XH-7RFJ] (last updated Nov. 26, 2024).

[62] Naomi Shiffman et al., Commentary, *Burden of Proof: Lessons Learned for Regulators from the Oversight Board's Implementation Work*, J. ONLINE TR. & SAFETY, Feb. 2024, at 3.

[63] *Oversight Board Upholds Meta's Decision in Altered Video of President Biden Case*, OVERSIGHT BOARD (Feb. 5, 2024), https://www.oversightboard.com/news/1068824731034762-oversight-board-upholds-meta-s-decision-in-altered-video-of-president-biden-case [https://perma.cc/JC5U-MAZX].

[64] Monika Bickert, *Our Approach to Labeling AI-Generated Content and Manipulated Media*, META (Apr. 5, 2024), https://about.fb.com/news/2024/04/metas-approach-to-labeling-ai-generated-content-and-manipulated-media [https://perma.cc/PDH2-WQ59].

[65] Shiffman et al., *supra* note 62.

[66] Shiffman et al., *supra* note 62, at 7.

The existence of regular reporting from both the Board and Meta on recommendations provides data that can be valuable for external audiences but should be taken with a grain of salt. As Evelyn Douek has pointed out, both Meta and the Board "have good reason to paint as glowing a picture of the Board's accomplishments as they can, in order to try convince outsiders of the benefits of the Board as an institution and reap the legitimacy dividends."[67]

The creation of the Oversight Board is a unique example of one company's investment in its own accountability mechanism, which from conception was oriented toward the challenging task of establishing global legitimacy for its own operational grievance mechanism. As extensively detailed by Klonick, the process that led to the Board entailed setting up a dedicated Governance Team within the company, a consultation with six global workshops, extensive additional meetings, and a public questionnaire.[68] This process informed the development of the Board's Charter, which sets out the relationship between Meta, the Board, and the Trust that funds the Board, to which the company has contributed $280 million in two tranches. Among the key decisions that Meta wrestled with in establishing the Board was deciding how Board members would initially be selected. Meta selected the first cadres of Board members who could in turn appoint future board members. The individuals appointed reflected a focus on global representation and a strong emphasis on internationally recognized experts on human rights and freedom of expression. Another measure of legitimacy for the Board is the extent of public involvement in the cases it adjudicates. The Board invites public comments to inform its deliberations, which are tracked and reported on in regular transparency reporting. The quality and quantity of these comments vary wildly depending on the notoriety of the issue in question. All of the procedural efforts at legitimacy, however, have yet to succeed in delivering other criteria, such as buy-in from other companies (a stated goal for the Board). Moreover, the sheer size of the financial commitment from Meta may keep public attention on the fact that the Board is funded entirely by the company it seeks to regulate. Because the Board is currently funded and focused entirely on one company that has become a lightning rod in the public debate around social media, it has also drawn critical

[67] Douek, *supra* note 60, at 405.

[68] Klonick, *supra* note 59, at 2454.

attention from activists who have dubbed themselves "the Real Facebook Oversight Board."[69]

As to accountability, the Oversight Board is acutely aware of its need to demonstrate impact. In addition to ensuring its binding judgments are upheld, this includes how its policy recommendations are considered and whether and how they are adopted. The development of teams and structures within the Board tasked with these issues attests to this strategic priority.[70] Whether these structures will survive layoffs,[71] persuade skeptics, and be sustained, will in part depend on the ability of the Board to consolidate. This could occur through adoption by other companies or interoperation with government regulations.[72]

C. Hash sharing to stop non-consensual intimate images

The non-consensual sharing of intimate images (NCII) is a particularly pernicious form of online abuse.[73] Amid efforts to criminalize this behavior in different jurisdictions across the United States and worldwide, companies have been working together with stakeholders and one another to find technical solutions to enable enforcement against NCII dissemination.

Stop Non-Consensual Intimate Image Abuse, or StopNCII.org, is a project operated by SWGfL, a UK nonprofit organization, to prevent the non-consensual sharing of intimate images through hash sharing. The project allows individuals to generate a digital fingerprint, or hash, of their intimate images.[74] Those hashes are

[69] *The Real Facebook Oversight Board*, THE CITIZENS, https://the-citizens.com/action/real-facebook-oversight-board [https://perma.cc/9JHY-H5VY] (last visited Nov. 9, 2024).

[70] Laurence R. Helfer et al., *The Meta Oversight Board's Human Rights Future*, 44 CARDOZO L. REV. 2233, 2273-2275

[71] Naomi Nix, *Meta's oversight body prepares to lay off workers*, WASH. POST (Apr. 29, 2024), https://www.washingtonpost.com/technology/2024/04/29/meta-oversight-board-layoffs [https://perma.cc/4CRQ-3TZN].

[72] For example, one pathway the Board is pursuing is certification as an out-of-court dispute settlement body as Article 21 of the EU's Digital Services Act. The Bord provided a grant to start Appeals Centre Europe, a new body certified by the Irish media regulator, Coimisiún na Meán, which says it will settle disputes relating to Facebook, TikTok, and YouTube. *See Resolving Content Disputes on Social Media*, APPEALS CENTRE EUROPE, https://www.appealscentre.eu/what-we-do [https://perma.cc/JQW6-X2GZ] (last visited Nov. 9, 2024).

[73] *See Trust & Safety Glossary of Terms*, DIGIT. TR. & SAFETY P'SHIP (July 2023), https://dtspartnership.org/wp-content/uploads/2023/07/DTSP_Trust-Safety-Glossary_July-2023.pdf [https://perma.cc/8VBX-V76V] (glossary entry for Non-Consensual Intimate Imagery (NCII)).

[74] According to the DTSP Trust & Safety Glossary of Terms, a hash function is an algorithm applied to inputs of variable length to provide a fixed length output. A given

shared with partner services who "look for matches to the hash and remove any matches within their system(s) if it violates their intimate image abuse policy."[75] Importantly, the actual images are not shared, just the hashes. Partners include social media, online dating, and adult services.

Interestingly, StopNCII started as a pilot collaboration between Facebook and Australia's eSafety Commissioner. This collaboration faced an initial wave of skeptical media coverage that focused on the role of Facebook employees viewing submitted nude photos as part of the pilot.[76] Widespread agreement among safety experts across industry and civil society provided momentum to expand this pilot via collaboration with nonprofit organizations in several countries. In December 2021, StopNCII.org was launched, using on-device hashing technology, so that the images do not have to leave the person's possession.[77]

Hash sharing is a relatively mature technical tool within the trust and safety toolbox that has been used over the past decades to address other forms of illegal or harmful content, particularly child sexual abuse material and, more recently, terrorist content. It is most effective with "exact matches of previously known violating content,"[78] which allows them to be defeated by manipulation of images via cropping, editing, and filtering. Notwithstanding technical improvements to increase the accuracy of fuzzy matching to defeat such efforts, there will always be instances of false positives and false negatives with any use of this technology.

From a legitimacy perspective, several components of StopNCII are worth noting. First, although initially developed by one company, StopNCII was housed at a nonprofit civil society organization with a long track record in the online safety ecosystem. Partnerships with dozens of nonprofits, victim advocates, and specialized regulators globally brought a degree of

hash algorithm (such as SHA-256) always returns the same value for a given input, making it a means of uniquely identifying a piece of digital content (such as an image, video, or block of text). *See Trust & Safety Glossary of Terms*, DIGIT. TR. & SAFETY P'SHIP (July 2023), [https://perma.cc/R34E-PZCA].

[75] *How StopNCII.org Works*, STOPNCII, https://stopncii.org/how-it-works [https://perma.cc/8U9V-WQ24] (last visited Nov. 9, 2024).

[76] *See* Brad Esposito, *Facebook Says Its Employees Will View Your Nudes If You Use Its Anti-Revenge Porn Program*, BUZZFEED (Nov. 9, 2017), https://www.buzzfeed.com/bradesposito/send-nudes-to-facebook [https://perma.cc/6VEC-2LVG].

[77] Antigone Davis, *Strengthening Our Efforts Against the Spread of Non-Consensual Intimate Images*, META, (Dec. 2, 2021), https://about.fb.com/news/2021/12/strengthening-efforts-against-spread-of-non-consensual-intimate-images [https://perma.cc/9F96-PZ4F].

[78] *FAQ The StopNCII.org tool*, STOPNCII, https://stopncii.org/faq [https://perma.cc/PH4B-3WBN] (last visited Nov. 9, 2024).

legitimacy that would not be possible with a company tool.[79] Hash-sharing collaborations are not without controversy, and in some cases have been accused of lacking adequate oversight.[80] A key distinction between StopNCII and other hash sharing efforts is that StopNCII only accepts hashes from the person depicted in the content, whereas with CSAM or terrorist content it is other people reporting this content, creating challenges around accountability.

In terms of accountability, the voluntary nature of self-submitting intimate image hashes to prevent NCII does not raise the same concerns, in volume or degree, as other cross-platform hash sharing efforts. Nonetheless, a process to examine how companies are using the tool would perhaps secure greater buy-in from stakeholders, including regulators, affected users, and the wider community. For example, the Child Sexual Abuse Material Hash List that is maintained by the National Center for Missing & Exploited Children (NCMEC) has been independently audited.[81] The Global Internet Forum to Counter Terrorism (GIFCT) conducted a human rights impact assessment, adopted a human rights policy, and regularly issues transparency reports[82] with more information about how its hash sharing database operates. These practices encourage accountability for cross-industry collaboration on enforcement against harmful content that nonetheless presents challenges with regard to freedom of expression and human rights. As the industry comes together to work across platforms in response to specific issues[83] and incidents,

[79] *See Global Network of Partners*, STOPNCII, https://stopncii.org/partners/global-network-of-partners [https://perma.cc/7C3H-VT6W] (last visited Nov. 9, 2024).

[80] *See* Evelyn Douek, *The Rise of Content Cartels*, KNIGHT FIRST AMENDMENT INSTITUTE (Feb. 11, 2020), https://knightcolumbia.org/content/the-rise-of-content-cartels [https://perma.cc/M4BM-BRM7].

[81] *See* Letter from David J. Slavinsky, Site Dir. of Concentrix, to John Shehan, Exploited Children Division & International Engagement Senior Vice President of Nat'l Center for Missing & Exploited Children (Apr. 12, 2024) (on file with the National Center for Missing & Exploited Children).

[82] *See* Global Internet Forum to Counter Terrorism, 2023 GIFCT Annual and Transparency Report (2023).

[83] Recent examples include: efforts to counter fake online reviews through the Coalition for Trusted Reviews, *see Amazon, Booking.com, Expedia Group, Glassdoor, Tripadvisor, and Trustpilot Launch First Global Coalition for Trusted Reviews* (October 17, 2023), https://press.aboutamazon.com/retail/2023/10/amazon-booking-com-expedia-group-glassdoor-tripadvisor-and-trustpilot-launch-first-global-coalition-for-trusted-reviews [https://perma.cc/GHP9-MLCX], as well as cross-industry collaboration on and "pig butchering" schemes, *see* Match Group, *Tech Companies Announce A New Coalition To Fight Online Fraud & Pig Butchering Scams* (May 21, 2024), https://ir.mtch.com/investor-relations/news-events/news-events/news-details/2024/Tech-Companies-Announce-A-New-Coalition-To-Fight-Online-Fraud—Pig-Butchering-Scams/default.aspx [https://perma.cc/M832-NHTP], to name just a few examples.

there will be an increasing need for consensus best practices to administer such efforts in line with trusted intermediary criteria.

D. Multistakeholder collaboration on risk assessment

One of the challenges of managing digital services that allow users to interact with each other, or to make and share content, is that they enable the full spectrum of human behavior. As a result, the risks to the rights and safety of users will evolve in unpredictable ways. Nonetheless, there is an evolving consensus across government, industry, and civil society that risk assessment methodologies should be a key component of digital safety.[84] More mature fields, such as information security, enterprise risk management, business, and human rights, all provide examples of risk management that have informed the development of digital safety risk assessments. More recently, online safety regulations enacted in Australia, the European Union, and the United Kingdom all specify some form of risk assessment for some digital products and services.[85]

At present, there is no agreed upon standard for what a digital safety risk assessment should entail. This creates challenges for services by having to prepare multiple risk assessments and follow different regulatory requirements for each of their services that are covered by these regulations, duplicating work in ways that may detract from overall trust and safety efforts. To encourage a more coherent approach, the World Economic Forum's Global Coalition for Digital Safety created a workstream on digital safety risk assessment,[86] which the author has co-chaired with the UK regulator Ofcom since spring 2022. In May 2023, the coalition published a report with a framework and set of case studies related to risk assessment. Other reports from the coalition include a set of principles for digital safety, a typology of online harms, and a report on measurements and metrics.

[84] As Ofcom wrote in their 2022 roadmap to regulation: "While the architecture of different national regimes might vary, the regulatory tools (e.g. transparency, risk assessment, audit) are likely to be common to most or all of them, and international cooperation, including through multistakeholder fora, can help us to develop a common regulatory toolkit informed by international best practice." Ofcom*, Online Safety Bill: Ofcom's roadmap to regulation* (July 6, 2022) https://www.ofcom.org.uk/siteassets/resources/documents/online-safety/information-for-industry/roadmap/online-safety-roadmap.pdf?v=328170 [https://perma.cc/X274-QCWQ].

[85] Mandatory risk assessments would face First Amendment concerns in the United States, especially in light of *NetChoice, LLC v. Bonta*, No. 23-2969, 2024 WL 3838423 (9th Cir. Aug. 16, 2024).

[86] *A Global Coalition for Digital Safety*, WORLD ECONOMIC FORUM, https://initiatives.weforum.org/global-coalition-for-digital-safety/home [https://perma.cc/9J28-7AY7] (last visited Nov. 9, 2024).

Assessing the effectiveness of this sort of coalition is not easy. Its stated goal is "to develop innovations and advance collaborations that tackle harmful content and conduct online."[87] Its track record of publications demonstrates an ability to do more than merely convene key stakeholders, but to release publications that represent at some level a consensus approach to thinking about complex and contested questions of online safety, even if the coalition members do not always agree on how to tackle these matters at a more granular level. The fact that the coalition has continued to add new members, without significantly losing members, as it publishes more reports, argues in favor of its effectiveness. However, whether its publications steer stakeholders toward real solutions is still to be seen.

The coalition's unique membership, which includes regulators like Ofcom and Australia's eSafety Commissioner, as well as other government representatives, companies, civil society organizations, and international organizations, provides a unique multistakeholder venue for trusted collaboration on common challenges. Although the World Economic Forum is sometimes viewed skeptically,[88] the coalition's wide membership does afford it an important level of legitimacy.

There is no accountability structure per se for participating in the coalition, nor are there requirements to use its outputs as part of risk assessments. Even so, the presence and active leadership in the coalition of regulators with statutory authority to mandate risks assessments is critical. As more information becomes public regarding regulatory risk assessments, analysis of whether and how both regulators and the regulated are using this sort of multistakeholder guidance will become possible.

E. Trust and Safety academic research

Support for academic research in the field of trust and safety is one of the highest profile areas of focus for legislators, regulators, companies, NGOs, and of course, academics themselves. Legislation proposing mandatory data sharing with vetted researchers has been proposed in the U.S. Congress,[89] and was

[87] *About Us A Global Coalition,* WORLD ECONOMIC FORUM, https://initiatives.weforum.org/global-coalition-for-digital-safety/about [https://perma.cc/4LP4-9K3A] (last visited Jan. 26, 2025).

[88] As Jean-Christoph Graz writes, "those closely associated with the Forum are inclined to deny its power and those fiercely opposed are likely to emphasize its overarching influence." Jean-Christophe Graz, *How Powerful are Transnational Elite Clubs? The Social Myth of the World Economic Forum*, 8 NEW POL. ECONOMY. 321, 321 (2003).

[89] *See* Platform Accountability and Transparency Act, S. 1876, 118th Cong. (2023).

enacted in the EU as Article 40 of the Digital Services Act.[90] Although companies have individually and collectively worked toward this objective, it is an area where company self-assessments have recognized there is substantial room for improvement.[91]

Data access raises complicated issues. How are researchers vetted? How can access to data be sufficient for research while protecting privacy and abiding by data protection regulations? Important progress on these matters is advancing particularly through the European Digital Media Observatory.[92]

At the same time that these developments have taken place over the last few years, a purely voluntary effort, that has not required any legislation or regulation, has substantially advanced the field of trust and safety research. At Stanford University, the inception of the Journal of Online Trust and Safety (JOTS) and the related Trust and Safety Research Conference (TSRC) have provided a critical new forum where academics and industry experts have shared path-breaking research, made concrete advances to understanding how risks and harms manifest, and shared what can be done about digital risks and harms.

JOTS has published seven issues since it debuted in October 2021, providing a place for both peer-reviewed research as well as wider commentary on trust and safety. The journal is complemented by the TSRC, which has provided an annual venue for sharing critical research, as well as building a community that crosses constituencies, allows for informal collaboration, and builds trust relationships between academics and practitioners. Academic research on matters related to trust and safety is not new, and it is worth noting that independent academics and researchers have often shed critical light on key challenges around issues such as algorithmic transparency and content moderation. However, JOTS and TSRC have significantly enhanced the research community, adding to the quantity and quality of scholarship in the field.

Although a full evaluation of the effectiveness of this work will require a retrospective analysis years in the future, JOTS tracks its own impact, noting many examples of attention in mainstream reporting, including having been featured in Federal Trade Commission reports and U.S. Senate testimony. In fact, the

[90] *See* Commission Regulation 2000/31, 2022 O.J. (277) 1.

[91] *See The Safe Assessments*, DIGIT. TR. & SAFETY P'SHIP (July 2022), https://dtspartnership.org/wp-content/uploads/2022/07/DTSP_Report_Safe_Assessments.pdf [https://perma.cc/LG7N-2TZY].

[92] *Access to Platform Data for Researchers*, EUROPEAN DIGIT. MEDIA OBSERVATORY, https://edmo.eu/areas-of-activities/research/access-to-platform-data-for-researchers [https://perma.cc/KN35-GAWN] (last visited Jan. 24, 2025).

proposed data sharing bill and the Platform Transparency and Accountability Act was shaped through commentary and model legislation published in the journal.[93] Contributions to the journal include a comprehensive look at how Zoom scaled their trust and safety operations during the COVID-19 pandemic[94] and an exploration of how Meta conducts stakeholder engagement on policy updates, among other topics.[95]

JOTS employs several mechanisms that support its legitimacy. These include a quick review system, its editorial board, and clear policies to address conflicts of interest. Maintaining the peer-reviewed section of the journal shows adherence to high standards of scholarship, which is further enhanced through a quick review system that allows publication decisions to be made at a faster rate than most scholarly journals. Its editorial board includes a multidisciplinary mix of leading experts on trust and safety, although it is weighted heavily toward North American experts and Stanford faculty. As researchers and practitioners increasingly recognize the importance of global majority perspectives in the field of trust and safety, it will be important for the journal to address this imbalance to help ensure global legitimacy in the future.[96]

As a research initiative that is independent of technology companies and financially supported by a private foundation, JOTS does not face the same questions that an industry effort would. Because it seeks to involve company practitioners, it must walk a careful line, with careful consideration of conflicts of interest. Ultimately, as more direct data access programs are implemented across digital services, these efforts will need to be accompanied by robust conformity assessments to ensure accurate and appropriate researcher access. But, importantly, JOTS shows that other purely voluntary, entrepreneurial research collaborations can adroitly bring new levels of insight into complex digital challenges in the absence of strict regulatory requirements.

[93] *See* Nathaniel Persily, *A Proposal for Researcher Access to Platform Data: The Platform Transparency and Accountability Act*, J. OF ONLINE TR. AND SAFETY, Oct. 2021, at 1.

[94] Karen Maxim, Josh Parecki & Chanel Cornett, *How to Build a Trust and Safety Team In a Year: A Practical Guide From Lessons Learned (So Far) At Zoom*, 1 J. OF ONLINE TRUST AND SAFETY, no. 4, July 2022, at 1, 4.

[95] Peter Stern, Sarah Shirazyan & Abby Fanlo, *How Can Platform Engagement with Academics and Civil Society Representatives Inform the Development of Content Policies? A Look at Meta's COVID-19 Misinformation Policies*, 1 J. OF ONLINE TRUST AND SAFETY, Napo. 4, Sept. 2022, at 1, 4.

[96] *Majority World Initiative*, INFORMATION SOCIETY PROJECT, https://law.yale.edu/isp/initiatives/majority-world-initiative [https://perma.cc/U8GJ-KW7Q] (last visited Nov. 4, 2024).

IV. ANALYSIS

These case studies represent a small sample of the very wide variety of industry and multistakeholder initiatives, partnerships, and other collaborations that have arisen in the past two decades to address online safety.

Skeptics will be quick to broadly paint such efforts as attempts to forestall regulation or manage the fallout from a crisis or scandal. The technology industry, which prides itself on operating at scale and with swift innovation, has faced a massive and rapid change in public perception. It has now become the preferred political punching bag of both the left and the right in Washington, D.C., state capitals, and around the world. A counterproductive consequence of this dynamic is that the kinds of regulatory innovation that are frequently championed in other industries have become politically unpalatable for digital services. Given the First Amendment restrictions on regulating harmful, but constitutionally protected expression in the United States, these sorts of voluntary collaborations are often the only way to approach such inherently expressive challenges.

What lessons can we learn from these case studies in how trusted intermediary criteria might enhance voluntary efforts to prevent and respond to harmful online content and behavior?

A. Assessing outcomes, not just outputs

Weiser calls for "a culture of retrospection" that rigorously evaluates what does and does not work in a given regulatory experiment, whether coming from the public or private sector.[97] In particular, he emphasizes a willingness to recognize failed experiments as part of this culture. In the examples explored in this article, retrospective analysis is often hampered by the nascency of these collaborations, the daunting societal change they seek to make, and the expectations of stakeholders about when and how that change should happen.

Drawing from the field of project management and monitoring and evaluation, one way to add rigor while acknowledging these challenges is to evaluate the effectiveness of initiatives at two levels: outputs and outcomes. Outputs refer to what an initiative did, and outcomes are what changed as a result of those activities.

At the level of outputs, it is possible to say that each of these initiatives are effective: content credentials are being deployed across hardware and software, the Oversight Board issues hundreds of recommendations to Meta, StopNCII supported more

[97] Weiser, *supra* note 6, at 2037, 2040.

than 570,000 images and videos, a global coalition agreed on a framework for assessing digital risks, and a journal is publishing dozens of influential articles on trust and safety each year.

However, expectations when it comes to online content and conduct are about outcomes rather than outputs. The long-term impact of any of these efforts will take years to evaluate, while the politics of these issues play out daily, if not hourly.

Technology companies pride themselves on being data-driven and having a culture of engineering. One way to demonstrate the credibility of industry commitments to effective experimentation is to share more results of rigorous impact assessment. This will require being forthright about what is working and what is not. Companies should ensure that dollars are devoted to impact analysis as part of a collaborative initiative they support. And they should be prepared to share more data with trusted external stakeholders. This will require additional investment at a time when budgets have broadly been cut across the industry.

B. Balancing legitimacy and adaptability

Formal structures and established mechanisms legitimize ad hoc entrepreneurial efforts but complicate how these initiatives can adapt and evolve. Complying with traditional norms around openness and transparency is considered a best practice for private regulatory initiatives and is an explicit requirement for formal standards development organizations.[98] Openness, however, is not always an unqualified characteristic in the adversarial realm of trust and safety, where transparency brings with it certain risks for practitioner organizations and individuals.[99] As a result, you find some innovative efforts evolving quietly, such as the initial efforts by Facebook to work with Australia's eSafety Commissioner to develop the hash matching approach for StopNCII, which might have been at risk of being undermined by bad actors had it been pursued via formal standards development. Quiet collaboration,

[98] The American National Standards Institute (ANSI) has created essential requirements for due process, the ANSI Essential Requirements. *ANSI Essential Requirements: Due Process Requirements for American National Standards,* AM. NAT'L STANDARDS INST. (Jan. 2024), https://share.ansi.org/Shared%20Documents/About%20ANSI/Current_Versions_Proc_Docs_for_Website/ER_Pro_current.pdf [https://perma.cc/2GMX-28RV].

[99] These risks include but are not limited to "safety risks for trust and safety teams resulting from their work, uncertainty around what might constitute improper coordination among companies, reluctance to educate adversarial bad actors about internal 'playbooks,' and a sense that revealing details about a particular user account being moderated failed to fully respect user privacy." Farzaneh Badiei, Alex Feerst & David Sullivan, Commentary, *Toward a Common Baseline Understanding of Trust and Safety Terminology*, 2 J. OF ONLINE TRUST AND SAFETY, no. 1, Sept. 2023, at 1, 2.

including among organizations and individuals who may viscerally disagree on other issues, can be a powerful means of building trust, but it is less clear that those mechanisms can be maintained over time in the absence of formal governance structures that are aligned with due process and transparency norms.

How much formality is required to gain legitimacy? And how might it affect the ability of an initiative to evolve and adapt to changing circumstances? Although we more commonly think of these structures as applicable to formal self-regulation, including certification regimes, the principles also apply in other areas. Clear, prominent, and consistently applied policies for disclosing conflicts of interest for JOTS authors, for example, provide transparency.

Given that all of the case studies concern initiatives that seek to address the harmful misuse of digital services, radical transparency is unlikely to provide a catchall solution, as too much transparency can provide guidance to those seeking to circumvent company policies against abusive content and behavior. Identifying credible partners who can serve as proxies for the public interest, allowing initiatives to evolve confidentially but with accountability, provides one path forward.

C. *Managing external and internal perceptions and expectations*

Even the most carefully designed initiatives can go awry in the face of perceptual friction once deployed externally. Unlike in other areas of law and technology, where the societal impact of policy may be obscured by layers of technical complexity, everyone who generates user content on digital services has a stake in how user-generated content is governed. Anyone who has had their content moderated, or engaged in moderation, likely has strong views about its governance, whether that takes the form of legislation and regulation or private sector initiatives. The complex and contested nature of services that facilitate expression tends to draw a far wider array of stakeholders and critics than more narrowly drawn efforts, evidenced by self-regulatory initiatives in other sectors or even more technically driven projects such as C2PA.

Gaps in how initiatives are designed and perceived can be broken down into two categories: external expectations and internal incentives. On the one hand, there are the expectations and perceptions of external audiences. With the Oversight Board, for example, years of preparation and hundreds of millions of dollars devoted to the project did not prevent criticism that the initiative remains controlled by the company that it is designed to hold accountable. Similarly, in the case of the Global Coalition for

Digital Safety, more broad perceptions of the strengths and weaknesses of the WEF as a whole and its association with Davos and a particular view of global politics and economics, may drive public perception of the coalition in ways that are far removed from its actual objective and activities.

How companies participating in these initiatives describe them is a second source of perceptual challenges. In the face of media scrutiny, companies may look to voluntary initiatives for validation. The risk is that this may overstate the immediate value of longer-term solutions, such as the applicability of content credentials and provenance standards as a means of mitigating immediate risks to electoral processes from AI and deepfakes. By precisely communicating the scope of collaborative projects and resisting the urge to overstate their aims, capacities, and impact, companies and other participants can preserve the credibility of these efforts over the longer term.

D. Rebalancing and future proofing through evaluation

These case studies argue in favor of heterogeneity when it comes to industry and multistakeholder initiatives to address online content risks. Diverse risks will require different constellations of expertise and influence that are not easily housed within any single institution, be it a multinational company or government agency. From technical standards to academic articles, diverse approaches can help ensure that companies, governments, and civil society do not focus too much on one particular approach to society-wide challenges. However, that does not mean that we cannot work toward more widely agreed approaches to evaluating the effectiveness of these efforts.

The risk assessments required by some regulators and informed by the WEF coalition work are primarily prospective and seek to inform the business decisions about deploying particular products or policies. There is also value, however, in aligning on methods for retrospective evaluation. DTSP has developed the Safe Framework as a method for assessing the maturity of a framework of industry best practices for trust and safety.[100] Although the case studies presented in this article broadly align with DTSP's five overarching commitments to product development, governance, enforcement, improvement, and transparency, not every initiative perfectly aligns with a specific best practice articulated by DTSP.[101]

[100] *The Safe Framework*, DIGITAL TRUST & SAFETY PARTNERSHIP 12 (Dec. 2021), https://dtspartnership.org/wp-content/uploads/2021/12/DTSP_Safe_Framework.pdf [https://perma.cc/PXW5-5LK3].

[101] *Id.* at 9.

The value of the Safe Framework is that it could conceivably be used to evaluate how any of these efforts are being used to address content and conduct-related risks, identify and document controls for those risks, and assess whether the controls are designed and operating effectively.

The broad adoption of this approach would help ensure a consistent approach to evaluating a diverse range of entrepreneurial self-regulatory efforts and provide confidence to decision makers in companies and regulatory agencies that their portfolio of policy approaches is balanced and appropriate.

CONCLUSION

The collaborative initiatives described in the article have emerged and operated within the legal framework of the First Amendment and Section 230, and as a result, proposed potential shifts in judicial interpretation of this framework loom large. In the past two years, the Supreme Court has taken up cases with enormous potential consequences regarding the scope of Section 230,[102] how the government interacts with social media companies,[103] and whether the First Amendment protects content moderation.[104] Importantly, the Court has resisted far-reaching reinterpretation of established precedent in each of these cases. By largely ruling on procedural matters, the Court has ushered in a new era of legal turbulence when it comes to user-generated content and the services that enable it.[105]

With the *Moody v. NetChoice* decision, six Justices have clarified that the First Amendment protects the rights of companies that provide expressive services to moderate the user-generated content they enable. This decision averts the short-term risk of conflicting state-level regulations that could have substantially interfered with the ability of companies to engage in the types of collaboration efforts described above. Further, in the longer-term, it sets a course toward clarifying the First Amendment rights of digital services in ways that will enable further entrepreneurial experimentation to address novel content challenges.

But, by rejecting the facial challenge brought by the plaintiffs and remanding the cases back to lower courts in Florida and Texas, the Court has opened the door to years of litigation. As Eric Goldman puts it, "the lower courts will need to consider how dozens

[102] Gonzalez v. Google, 598 U.S. 617, 622 (2023).

[103] Murthy v. Missouri, 144 U.S. 1972, 1983 (2024).

[104] Moody v. NetChoice, 144 U.S. 2383, 2388 (2024).

[105] *See generally* Evelyn Douek & Genevieve Laiker, *Lochner.com?*, 138 HARV. L. REV. 100 (2024).

of statutory provisions could apply to dozens of potentially regulated entities that each have multiple communication modalities—a daunting multi-dimensional project for all involved."[106] And that is just for these two particular statutes. In the meantime, the wave of state bills that have been enacted and are facing litigation continues to grow. State lawmakers may now be incentivized to pass even more sweeping laws, applied to a broader set of services, which would raise the costs associated with litigation.[107] As a result, it will be years before we have a clear picture as to what kinds of regulations around the governance of online speech and behavior will be durable.

Returning to the portfolio approach analogy, prudent financial planners do not exit the market during periods of volatility. Instead, they double down on diversifying assets, diligently managing risk, and looking for long term opportunities. A similar approach should guide decision makers with a stake in the governance of digital products and services. Collaborative experimentation within industry, and with other key stakeholders in academia, civil society, and government will provide opportunities for agile prevention and response to acute risks of online abuse. More rigorous application of criteria for trusted intermediaries can ensure that investment and energy go into those initiatives that are best structured for success. Not every initiative can or should succeed, but those that can measure success in terms of concrete improvements to the safety and rights of internet users have the potential to play as important a role in the future of digital governance as any law, policy, or judicial decision.

[106] Eric Goldman, *"Speech Nirvanas" on the Internet: An Analysis of the U.S. Supreme Court's Moody v. NetChoice Decision*, 23 CATO SUP. CT. REV., at 125, 137 (2024).

[107] Jess Miers, *The Messy Reality Behind Trying To Protect The Internet From Terrible Laws*, TECHDIRT (July 25, 2024), [https://perma.cc/AF5V-FAMZ].