Commentary

Toward a Common Baseline Understanding of Trust and Safety Terminology

Farzaneh Badiei, Alex Feerst, and David Sullivan

1 Introduction

As a field of study, trust and safety is relatively new. The operations and processes that it includes, however, have been in place for many years, dating back to the beginning of networked communications and becoming an increasingly common function within companies as the commercial internet developed. Unlike other fields within the wider world of technology, such as cybersecurity or privacy, trust and safety has lacked an accepted and authoritative glossary of key terms. Glossaries can help with clarifying concepts and definitions. For example, the introduction to the National Institute of Standards and Technology (NIST) Glossary of Key Information Security Terms states: "Over time, use of this Glossary will help standardize terms and definitions used, reducing confusion and the tendency to create unique definitions for different situations" (NIST 2019).

The Digital Trust & Safety Partnership (DTSP) brings together providers of diverse digital products and services around a shared framework of best practices for trust and safety (DTSP 2021).[1] As part of our objective of articulating and promoting best practices, we have developed a Trust & Safety Glossary of Terms that professionals use in their day-to-day operations (Digital Trust and Safety Partnership 2023). 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.

This article unpacks our descriptive approach to developing a trust and safety glossary. We explain the motivations behind assembling a glossary, our method and its limitations, and the challenges we encountered when putting the glossary together. We also discuss how we will proceed from here, as we aim to provide an evolving, increasingly comprehensive but easy to use glossary for different stakeholders.

2 Why a glossary?

Controversy around undesirable online content and conduct, from online harassment to copyright infringement, is not a new issue (ACLU Cyber Liberties Update 1995). In the early days of the internet, trust and safety operations, often conducted by network administrators or volunteers, faced many of these same issues (Dibbell 1994).

As a professional function within digital services, trust and safety developed quietly over the course of several decades (Cryst et al. 2021). Since roughly 2004, the rise of social

[1] DTSP partner companies are listed at https://dtspartnership.org/.

doi:10.54501/jots.v2i1.147

media and other Web 2.0 services led to content policies being developed and enforced centrally, often with increasing use of automated content moderation systems (Bozdag 2013).

Throughout this period, how trust and safety was provided on various platforms (particularly, the internal rules for content policy, and how enforcement operations worked) remained largely confidential, with reasons including 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 (Feerst 2019).[2] Rather, details about trust and safety became public through work by journalists, advocates, and scholars. While it is not possible to include an exhaustive list of such references, we have included a number of particularly important illustrative examples of journalistic reporting on content moderation (Chen 2014; Newton 2019), books about advocacy on trust and safety issues (MacKinnon 2012; York 2021), and academic articles and books (Baym and boyd 2012; Klonick 2017; Gillespie 2018). A major milestone for transparency was the first Content Moderation at Scale Conference in 2018, where trust and safety practitioners publicly and voluntarily shared facts and data about their work (SCU 2018).

A downside to the confidential approach to trust and safety was that it was hard to share knowledge, both for short-term tactical purposes (like responding to cross-platform campaigns) and longer-term education and preventive planning. The Trust and Safety Professional Association (TSPA) has published a trust and safety curriculum, which succinctly describes the challenge: "In the past, that hard-won knowledge was shared through one-on-one mentoring or coffee-meet-ups and difficult to share widely" (TSPA 2021).

In recent years, the rise of new institutions including DTSP, TSPA, this journal, and others have begun to change the inward-facing nature of trust and safety, presenting a professional field that is taking steps toward maturity. But practices that developed in silos within companies that were often reluctant to share too much information created gaps in understanding, which were in turn magnified by trust and safety becoming the focus of public interest and attention from policymakers, as well as deficits in trust among companies and various stakeholders, such as journalists and civil society.

In the politically charged conversations happening around the world and at all levels of government about how online content and conduct should be regulated, some shared vocabulary is essential. Research on other highly charged situations of conflict has shown that glossaries can improve trust between stakeholders and help facilitate policy discussions and negotiations (*Statement by the Permanent Members of the Security Council (P5) of the United Nations at the 2015 Review Conference of the Parties to the Treaty on the Non-Proliferation of Nuclear Weapons (NPT)* 2015).[3] With this in mind, we began our glossary development process.

[2] In interviews with trust and safety practitioners, Alex Feerst noted, "most use pseudonyms to correspond with users because they've been threatened or stalked online."

[3] In 2009 the United Kingdom started a process with the Parties to the Treaty on the Non-Proliferation of Nuclear Weapons (NPT) to create a glossary. One goal of the glossary was to be reviewed and improved by civil society participation to create trust, transparency, and understanding of the processes: "We continue to implement Action 5 of the Action Plan to 'further enhance transparency and increase mutual confidence' through P5 dialogue and action. In this regard, we agreed on a common reporting framework in 2014 under France's leadership and completed a first edition of a Glossary of Key Nuclear Terms under China's leadership."

3 Our objectives and approach

A glossary of terms was one of the first projects raised among members following DTSP's launch in February 2021. Work began in earnest in early 2022. Our objectives were to:

- **Aid the professionalization of the field and support nascent trust and safety teams as they build out their operations**. We wanted the glossary to be a useful companion resource for a new hire working in trust and safety operations, content policy, or other relevant functions.
- **Support the codification of agreed interpretations of critical terms used across the industry**. Although precise definitions are likely to vary based on specific products, audiences, and jurisdictions, baseline definitions could help companies across the industry sing from the same song sheet.
- **Facilitate informed dialogue between industry, policymakers, regulators, and the wider public**. Growing interest in how trust and safety works has not yet led to improved awareness of what the job looks like inside companies. By increasing transparency around the meaning of key terms, our intention is to bridge the gap between practice, policy, and the wider political context in which trust and safety is debated. Within this objective, we also saw particular value in enhancing scholars' understanding of trust and safety terminology, to improve their analysis of the governance and operations of digital services.

4 Assessing the landscape

To carry out this project, we first looked to existing efforts across industry, civil society, and governmental processes.

We identified a number of similar efforts to define trust and safety-related terminology. Existing industry glossaries include the TSPA Glossary for the Trust and Safety Curriculum, and glossaries developed by vendor companies such as ActiveFence (ActiveFence 2023) and Tremau (Tremau 2023).

Outside of industry, academics and civil society use concepts such as "platform accountability," "platform governance," "responsible tech," and "ethical tech" to articulate the concerns that most companies would address through a trust and safety lens to refer to many of the same issues covered by trust and safety (Task Force for a Trustworthy Future Web 2023). A multistakeholder process developed by the Internet Governance Forum Coalition on Platform Responsibility developed a Glossary of Platform Law and Policy Terms presented at the 2021 Internet Governance Forum (Belli, Zingales, and Curzi 2021). While this highly detailed publication is an impressive outcome from a multistakeholder process, its sheer length presents challenges both for useability and for its ability to be regularly updated.[4]

We also took note of regulatory and civil society developments relevant to a trust and safety glossary. These include the many legislative efforts to regulate online content and services around the world,[5] and increased coordination between online safety

[4] Notably, this publication does not include a definition for "trust and safety."

[5] Recent examples of content regulations include Australia's Online Safety Act of 2021, the European Union's Digital Services Act, which came into force in November 2022, and the UK Online Safety Bill, which is expected to be enacted in Fall 2023.

regulators.[6]

There are also other normative principles put forward by civil society and non-state actors. Civil society organizations have also played a critical role in issuing principles that call for transparency and accountability from companies, such as the Manila Principles on Intermediary Liability ("Manila Principles for Intermediary Liability" 2015) and the Santa Clara Principles on Transparency and Accountability in Content Moderation ("Santa Clara Principles on Transparency and Accountability in Content Moderation" 2018). These initiatives tend to be rooted in international human rights law and have broad operational principles about disclosure and transparency, and are not issue specific.

Finally, international organizations are increasingly involved in matters related to online trust and safety. This includes human rights-oriented initiatives such as the B-Tech Project at UN Human Rights (High Commissioner for Human Rights 2023), as well as UNESCO's proposed guidelines for regulating digital platforms (UNESCO 2023).

5 Methodology

To create the glossary, we combined individual expertise with broad industry and stakeholder consultation and review. As editor-in-chief for this project, Alex Feerst brought his background as a trust and safety practitioner and combined it with formal training in both law and English literature. To validate the definitions and enhance accuracy, we interviewed industry leaders and gathered feedback and commentary from trust and safety experts. As part of this process, DTSP partner companies reviewed the definitions. The first version of the glossary was available for public preview in relevant trust and safety conferences such as TrustCon2022.[7] We then opened a call for public comments by the general public, and DTSP members also were given a chance again to provide their feedback. We received comments from regulators, academics, and civil society organizations, worldwide.

6 Limitations

We approached our goal aware of certain limitations inherent in undertaking a glossary in the English language and with an industry-led approach.

Regarding language, we acknowledged that an English-first approach to trust and safety may reinforce the unequal role of English compared with other languages when it comes to content moderation and trust and safety. Studies have demonstrated that human and automated moderation in other languages lags substantially behind English (CDT 2017). Nonetheless, given the global influence of US-based digital products and services, for whom English is the lingua franca of their daily operations, for our descriptive effort it was necessary to start with English. Our goal is for this glossary to contribute to more fruitful and informed discussions about how the field of trust and safety can evolve to reflect the lived experience of those people who use and are affected by digital services all around the world. As the Task Force for a Trustworthy Future Web concluded, "companies' internal systems are often not tailored for the needs of partners from the Majority World, and not enough has been done to engage such partners proactively in anticipating the evolution of local risk factors, harms, and user needs" (Task Force for a Trustworthy

[6] See The Global Online Safety Regulators Network that launched in November 2022: https://www.esafety.gov.au/about-us/who-we-are/international-engagement/the-global-online-safety-regulators-network.

[7] TrustCon is an annual trust and safety conference bringing trust and safety professionals, regulators, and civil society actors together. See more at https://www.trustcon.net/.

Future Web 2023).

Our decision to lead with industry understanding of terms, rather than what a term ought to mean, was deliberate. Although a wide array of stakeholders, from policymakers to the individuals who use or are affected by a given digital service, may have important views on what terms should mean, we took a more targeted approach. Getting industry to coalesce around baseline definitions is a necessary first step toward a more nuanced discussion of expectations for online trust and safety.

7 Key issues that arose

Through the process of drafting, reviewing, and incorporating comments, the following key issues arose.

7.1 Usability, consensus, and commentary

The consultation draft of our glossary aimed to have short definitions of a few sentences, easily scanned and understood by a busy, recent entry-level hire into the field. But the process of development, preconsultation, caused many definitions to swell into lengthy paragraphs, often including examples of what the term "may" refer to or other examples of context that varied depending on how the term was used by a given product, related to laws in a particular jurisdiction, and other details.

Following the public consultation, we opted to employ a new structure in order to improve brevity and readability while preserving the option to add greater context and examples to definitions. The definitions themselves were limited to only one or two sentences, and we added a "commentary" section with additional bulleted information, much of which was moved out of the consultation draft definitions.

By shortening the definitions, we also increased the degree of consensus around the core baseline definition of key terms. And we can include additional detail in the commentary section in future iterations of the glossary, providing a forum in which diverse viewpoints can be ventilated about how particular terms are understood within the industry.

7.2 Developing a descriptive output amid normative expectations

As stated above, our objective was to define terms as they are used by practitioners within the field, so that clear communication of current practice can provide a foundation for more productive future discussion.

This is a different approach from thinking about how terms *should* be defined, according to the viewpoints of other stakeholders. But if we do not understand or clearly communicate current practice, efforts to improve that practice will run aground.

An example of this tension is around how trust and safety practices relate to the human rights responsibilities of digital companies. There has been a longstanding effort by human rights organizations and experts in academia, civil society, and inside companies to align company content moderation with international human rights law (Kaye 2019). While companies have developed dedicated human rights teams and policies, whose work intersects and overlaps with those of trust and safety teams, it was important in our view to be clear that trust and safety functions often predated the development and application of business and human rights to digital companies.

Looking ahead, it is clear that there is substantial overlap and opportunities for trust and safety and human rights teams to learn from each other, and we expect our definitions to evolve as this process takes place.

7.3 Drawing boundaries for an inherently interdisciplinary field

Trust and safety is a fundamentally interdisciplinary field that consists heavily of "cross-functional teams" (TSPA 2021). Safety issues are often also privacy or security issues, and there is a need for these fields to work together, inside and outside of companies (Polgar 2022).[8] For example, the DTSP Safe Assessments report found that "different teams and functions, including engineering and marketing, bring unique perspectives and may use different terms to describe common challenges. In some cases, the assessments also generated insights for other functions, such as privacy or security" (Digital Trust and Safety Partnership 2022).

In early drafts, we included terms such as "Data Protection," a concept that is often associated with the privacy profession. In this case, the International Association of Privacy Professionals (IAPP) maintains a comprehensive glossary that is used for its professional certifications, and so we decided to remove terms that we saw as owned by this community and resource (IAPP 2023).

We also included terms such as "Denial of Service/Distributed Denial of Service," which is commonly considered an information security issue. Numerous glossaries of information security terms exist, including the NIST Glossary of Key Information Security Terms, and we opted to remove this term and not include similar information security terms, as they were well defined elsewhere (NIST 2019).

There undoubtedly remain terms in our glossary that are defined elsewhere and may be defined differently by other related fields, including but not limited to privacy and security. Should discussions ensue regarding the differences between how trust and safety defines these terms compared with other groups, we consider that a positive outcome from this exercise.

7.4 An academic discipline or a professional field?

In a thoughtful reply to the public consultation, experts at noted Argentine academic center for the study of freedom of expression El Centro de Estudios en Libertad de Expresión y Acceso a la Información (CELE) wrote:

> *Trust and Safety does not seem to be a discipline in the academic sense of the word: it is, rather, a way of grouping certain practices within private corporations who work as intermediaries in the information flow that happens through the Internet. We are all for it, but we would like to make the point: defining it as a discipline may project the wrong idea that the field is crossed by broad consensus on a set of shared concerns, the problems that must be addressed, and so on. We are not sure that is the case.* (CELE 2023)

Whether trust and safety should be considered an academic discipline is outside the scope of our glossary. The Journal of Online Trust and Safety and the teaching curriculum developed by Stanford and other partners point to the emergence of research and

[8] David Ryan Polgar describes this dynamic: "After interviewing dozens of people working in Trust and Safety for projects and reports related to All Tech Is Human, it has become apparent to me that those in the field have an ability to work cross-functionally, traversing multiple departments. Often, these Trust and Safety professionals pointed out a 'non-linear career path' that led them into the field. I realized that a non-linear career path, one that was multi-sector and often multidisciplinary, was less of an outlier and more of a clear advantage for individuals who were now being rewarded for their broad backgrounds. Reading over dozens of job descriptions in Trust and Safety, a clear trend is that tech companies are looking for people who are comfortable engaging with multiple departments."

teaching dedicated to trust and safety, but that has happened disproportionately in the United States, less so in the Majority World (SIO 2023).

For DTSP, our aim has been to articulate the practices used by our members to keep services safe and trustworthy, and to define the key terms and concepts that are part of those practices. This effort is deliberately industry-led, with the intention that our work product forms a contribution to a growing, global conversation about how to handle risks and safeguard rights online.

That said, we recognize the lack of a broad consensus on many of the concerns touched upon by trust and safety and the terms we have defined. This is especially true from a global perspective (see more on this below).

Our aim with this glossary was both to present the understanding of terms used within industry, and to facilitate greater reflection on those terms from across stakeholder groups and around the world.

7.5 Tensions between legal and functional definitions

Depending on the structure of a given company, trust and safety can fall inside or outside of the legal department. And while many trust and safety professionals are lawyers, others are engineers, social workers, former government or law enforcement professionals, or customer experience professionals.

Evaluating and complying with a host of laws is a key part of the trust and safety field, but terms that relate to the law may have a different understood, functional meaning within trust and safety. For example, copyright has a specific legal meaning in the US legal context, a meaning that is in tension with other understandings of intellectual property concepts and laws in other jurisdictions. But within trust and safety, copyright is functionally understood as the set of laws that give rise to takedown notices for allegedly infringing content, with which trust and safety teams must contend, often at scale.

In thoughtful comments, the Samuelson-Glushko Canadian Internet Policy and Public Interest Clinic (CIPPIC) raised concerns that our proposed definition “seems to prioritize the interests of rightsholders over those of the public” (CIPPIC 2023). We adopted many of the recommended changes proposed by CIPPIC, and we opted to include commentary that noted the influence of US copyright laws on the practice of trust and safety, in the spirit of describing the functional use of such terms by practitioners.

We also considered whether it was necessary to include terms specific to intellectual property law that could be viewed as standard legal terms. We concluded that in the spirit of describing the terms used by practitioners in the field as it currently exists, intellectual property considerations, and specifically the intellectual property laws of the United States, are understood and play a sufficiently significant role in the field of trust and safety that they merited inclusion. Nonetheless, we emphasized at the outset that we are not providing legal definitions in the glossary.

7.6 Avoiding a US-centric approach while describing a US-dominated field

We were not surprised to receive several comments that noted our public consultation draft was tilted toward the United States.[9] Nearly everyone involved in the Glossary’s development either lives in the United States or works for a US-headquartered company.

[9] For example, see the following comment from the Australian eSafety Commissioner: “Many of the terms and definitions are very U.S.-centric (for example, references to Section 512 of the Copyright Act 17 U.S.C. § 512], ‘Safe Harbor’, and the definition of ‘Defamation’). Further international examples should be used to ensure the terms are reflective of the global reach of online platforms.”

The comments we received from individuals and organizations from outside the United States helped us identify implicit bias in our approach and brought additional rigor to revisions.

In particular, we decided to remove references to the United States, US law, and references to other specific jurisdictions from the definitions. We also limited such references in the commentary to only a few cases where we felt they were necessary. For copyright-related terms, because of the extent to which US-based companies are subject to takedown requests under the Digital Millennium Copyright Act, which can result in content being removed globally, our view was it was important to keep these references.

Elsewhere, we added references to other jurisdictions to show the broader range of legal approaches that wind up touching upon issues within the remit of trust and safety teams. For example, we broadened the definition of defamation to go beyond that within US law and noted in the commentary that "differences in legal approach and levels of associated legal risk may influence the takedown processes for defamation disputes adopted by online services in various localities."

7.7 Defining content-specific terms as a content-agnostic initiative

DTSP is a content-agnostic approach to maturing the field of trust and safety. Our partnership has identified a framework of best practices that companies can use to manage their content- and conduct-related risks, but we do not seek to align companies around common content policies or terms of service.

So why are we in the business of defining content-specific terms like "hate speech," "glorification of violence," and other heavily contested and controversial terms?

After consideration and consultation, we saw value in articulating short definitions that offer a baseline understanding of these terms, which are at the center of public debate about digital products and services. Specific definitions for prohibited content or conduct will always need to be product-specific, because the purpose and functionality of a search engine is substantially different from a social network. Similarly, an online game may provide a very different context for certain behavior compared to a service that combines online and offline behavior such as a dating application or transportation service. Content policies will also differ depending on the values, cultures, and brands that providers bring to their products.

We also saw value in pointing out where there is a lack of international consensus on legal definitions for terms such as hate speech and terrorism, given the particular challenges for globally operating companies who must contend with legally prohibited content that varies considerably across jurisdictions. We drew from authoritative international efforts to align terminology through expert work, such as the Luxembourg Principles, terminology guidelines for the protection of children from sexual exploitation and sexual abuse (Greijer and Doek 2016). Another valuable issue-specific initiative is the work done by the Global Internet Forum to Counter Terrorism (GIFCT) and its partners on legal frameworks for terrorist and violent extremist content (GIFCT 2023).

These and other initiatives inside and outside industry will be where the most in-depth discussions on these issues should take place, raising the ceiling. Our objective is more limited, improving understanding across trust and safety practitioners to raise the floor.

8 Where we go from here

A glossary that sits on a shelf is of little use to the emerging trust and safety profession. Our goal was to create a document that would be useful inside and outside of industry, helping to bridge the gaps in understanding that have created obstacles to a global conversation about what trust and safety should look like. To that end, our goal is not only to share the glossary, but also to articulate these reflections on its development, with as many interested parties as possible.

Provided that the glossary demonstrates value to our partner companies and the larger digital industry, we intend for this document to evolve over time, with the development of updates providing a place for thoughtful deliberation about what terms define the future of trust and safety.

References

ACLU Cyber Liberties Update. 1995. *AOL Censors Gay Video Titles, Finds "Buns" Acceptable but "Studs" Too Sleazy.* https://archive.epic.org/free_speech/censorship/aol_censorship.txt.

ActiveFence. 2023. *The Trust and Safety Glossary.* Glossary. ActiveFence. https://www.activefence.com/the-trust-safety-glossary.

Baym, Nancy K., and danah boyd. 2012. "Socially mediated publicness: An introduction." *Journal of Broadcasting & Electronic Media* 56, no. 3 (September 11, 2012): 320–29. https://doi.org/10.1080/08838151.2012.705200.

Belli, Luca, Nicolo Zingales, and Yasmin Curzi. 2021. "Glossary of Platform Law and Policy Terms," December. https://www.intgovforum.org/en/filedepot_download/45/20436.

Bozdag, Engin. 2013. "Bias in Algorithmic Filtering and Personalization." *Ethics and Information Technology* 15 (June 23, 2013): 209–27. https://doi.org/10.1007/s10676-013-9321-6.

Center for Democracy and Technology. 2017. *Limits of Automated Social Media Content Analysis.* Technical report. CDT. https://cdt.org/wp-content/uploads/2017/11/2017-11-13-Mixed-Messages-Paper.pdf.

Centro de Estudios en Libertad de Expresión y Acceso a la Información. 2023. "Glossary of Trust & Safety Terms, CELE's Submission." https://www.palermo.edu/Archivos_content/2023/cele/papers/CELE_submission_Glossary.pdf.

Chen, Adrian. 2014. "The Laborers Who Keep Dick Pics and Beheadings Out of Your Facebook Feed." *Wired* (October 23, 2014). https://www.wired.com/2014/10/content-moderation.

Cryst, Elena, Shelby Grossman, Jeff Hancock, Alex Stamos, and David Thiel. 2021. "Introducing the Journal of Online Trust and Safety." *Journal of Online Trust and Safety* 1 (1). https://tsjournal.org/index.php/jots/article/view/8.

Dibbell, Julian. 1994. "A Rape in Cyberspace." In *Flame Wars: The Discourse of Cyberculture,* edited by Mark Dery, 237–61. Duke University Press, January. ISBN: 978-0-8223-9676-5. https://doi.org/10.1215/9780822396765-012.

Digital Trust and Safety Partnership. 2021. *DTSP Best Practices Framework.* Best Practices. Digital Trust and Safety Partnership. https://dtsp.wpengine.com/wp-content/uploads/2021/04/DTSP_Best_Practices.pdf.

———. 2022. *DTSP Safe Assessments Report.* Report. Digital Trust and Safety Partnership, July. https://dtspartnership.org/dtsp-safe-assessments-report.

———. 2023. *DTSP Trust and Safety Glossary of Terms.* Glossary. DTSP, July. https://dtspartnership.org/wp-content/uploads/2023/07/DTSP_Trust-Safety-Glossary_July-2023.pdf.

Feerst, Alex. 2019. "Your Speech, Their Rules: Meet the People Who Guard the Internet." OneZero. https://onezero.medium.com/your-speech-their-rules-meet-the-people-who-guard-the-internet-ab58fe6b9231.

Gillespie, Tarleton. 2018. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions that Shape Social Media.* Yale University Press. ISBN: 978-0-300-23502-9. https://doi.org/10.12987/9780300235029.

Global Internet Forum to Counter Terrorism. 2023. "The Definitions and Principles Framework Project." https://def-frameworks.gifct.org.

Greijer, Susanna, and Jaap Doek. 2016. *Terminology Guidelines for the Protection of Children From Sexual Exploitation and Sexual Abuse.* Interagency Working Group on Sexual Exploitation of Children, June. ISBN: 978-92-61-21501-9. https://ecpat.org/wp-content/uploads/2021/05/Terminology-guidelines-396922-EN-1.pdf.

High Commissioner for Human Rights, United Nations Office of the. 2023. "B-Tech Project." https://www.ohchr.org/en/business-and-human-rights/b-tech-project.

International Association of Privacy Professionals. 2023. "Glossary of Privacy Terms." https://iapp.org/resources/glossary/.

Kaye, David. 2019. *Speech Police: The Global Struggle to Govern the Internet.* Columbia Global Reports. ISBN: 978-0-9997454-8-9. https://doi.org/10.2307/j.ctv1fx4h8v.

Klonick, Kate. 2017. "The New Governors: The People, Rules, and Processes Governing Online Speech." *Harvard Law Review* (March 21, 2017). https://ssrn.com/abstract=2937985.

MacKinnon, Rebecca. 2012. *Consent of the Networked: The Worldwide Struggle for Internet Freedom.* Basic Books, January 31, 2012. ISBN: 978-0-465-02442-1.

"Manila Principles for Intermediary Liability." 2015. https://manilaprinciples.org.

National Institute of Standards and Technology. 2019. *Glossary of Key Information Security Terms.* NISTIR 7298 Revision 3. National Institute of Standards and Technology. https://doi.org/10.6028/NIST.IR.7298r3.

Newton, Casey. 2019. "The Trauma Floor: The Secret Lives of Facebook Moderators in America." *The Verge* (February 25, 2019). https : //www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona.

Polgar, David Ryan. 2022. "How Do You Launch a Career in Trust and Safety?," May 3, 2022. https://builtin.com/career-development/launch-career-trust-safety.

Samuelson-Glushko Canadian Internet Policy and Public Interest Clinic. 2023. "Consultation on Trust & Safety Glossary of Terms – Definition of Copyright." https://dtspartnership.org/wp-content/uploads/2023/08/Consultation-on-Trust-Safety-Glossary-of-Terms-%E2%80%93-CIPPIC-Submission.pdf.

"Santa Clara Principles on Transparency and Accountability in Content Moderation." 2018. https://santaclaraprinciples.org.

Santa Clara University School of Law. 2018. *Content Moderation and Removal at Scale.* Conference, February 2, 2018. https://law.scu.edu/event/content-moderation-removal-at-scale.

Stanford Internet Observatory. 2023. "The Trust & Safety Teaching Consortium." https://stanfordio.github.io/TeachingTrustSafety.

Statement by the Permanent Members of the Security Council (P5) of the United Nations at the 2015 Review Conference of the Parties to the Treaty on the Non-Proliferation of Nuclear Weapons (NPT). 2015. https://www.un.org/en/conf/npt/2015/statements/pdf/P5_en.pdf.

Task Force for a Trustworthy Future Web. 2023. *Scaling Trust on the Web.* Atlantic Council, June. ISBN: 978-1-61977-279-3. https://www.atlanticcouncil.org/in-depth-research-reports/report/scaling-trust.

Tremau. 2023. “Trust & Safety Glossary.” https://tremau.com/glossary.

Trust & Safety Professional Association. 2021. “Introducing the Trust and Safety Curriculum.” https://www.tspa.org/2021/06/17/introducing-the-trust-and-safety-curriculum.

UNESCO. 2023. “Internet for Trust - Towards Guidelines for Regulating Digital Platforms for Information as a Public Good, Paris, 2023.” https://unesdoc.unesco.org/ark:/48223/pf0000384031.

York, Jillian C. 2021. *Silicon Values: The Future of Free Speech Under Surveillance Capitalism.* Verso Books. ISBN: 978-1-78873-880-4. https://www.versobooks.com/en-gb/products/882-silicon-values.

Authors

Farzaneh Badiei is the head of outreach and engagement at Digital Trust and Safety Partnership. She is the founder of Digital Medusa, an initiative that focuses on protecting the core values of our global digital space with sound governance.

Alex Feerst is a co-founder and advisor to the Digital Trust and Safety Partnership. He leads Murmuration Labs, which helps tech companies address and mitigate the human and social risks of innovative products.

David Sullivan is the founding Executive Director of the Digital Trust and Safety Partnership.

(dsullivan@dtspartnership.org)

Acknowledgements

We thank the partner companies of the Digital Trust and Safety Partnership for their contributions to the glossary. We also greatly appreciate the valuable input received from academic organizations, industry partners, regulators, and other global stakeholders during the public consultation held earlier this year. DTSP is funded by its partner companies; a current list can be found on its website at dtspartnership.org.

Keywords

Trust and safety, glossary, best practices.