Combating Child Sex Trafficking Material Online

A Study of Meta's, MindGeek's, and Snap Inc.'s Transparency Reporting

Written By:

Emelie Beckman

Advised By:

Professor Tracey Holland

HRTS UN3994 Human Rights Thesis Seminar - Fall 2024

Institute for the Study of Human Rights

Columbia University in the City of New York

Key Words: content moderation, child sex trafficking material, human rights of children, transparency.

ABSTRACT

Combating Child Sex Trafficking Material Online

A Study of Meta's, MindGeek's, and Snap Inc.'s Transparency Reporting

Emelie Beckman

The increasing spread of child sex trafficking material (CSTM) online is a violation of children's rights. The U.S. government requires companies that own social media platforms where perpetrators distribute material to monitor their platforms to prevent the spread of CSTM. In recent years, these companies have begun publishing transparency reports detailing how they moderate this harmful content. However, researchers criticize the transparency reports for not being transparent. This thesis concludes that two of these companies, Snapchat and Facebook, consistently fail to disclose data regarding their automatic moderation systems in these reports, the primary tool for identifying CSTM online. Although they have increased the amount of information in their reports, they have yet to include meaningful data regarding their content moderation of CSTM. Therefore, they have not increased their transparency regarding what actions they take against CSTM and have failed to protect children's rights.

Table of Contents:

List of Figures:

AKNOWLAGMENTS

I would like to extend my gratitude to the many people who have helped me throughout my research process and in the writing of this thesis. Firstly, I would like to thank my classmates, Maria-Vera Azadina, Anabella Dovale, Bokyoung Jeon, Clarissa Lara, Ruth Meschery-McCormack, Andrada Nicolae, Laura Romero Espinoza, Lara Selcuker, and Jackson Stanley. Your research projects have been truly inspirational, and your feedback has been invaluable to the successful completion of my research.

Secondly, I want to thank my friends and family for their involvement throughout this process and for cheering me on during the many ups and downs of writing this thesis. Without your patience, support, and encouragement, I would not have been able to complete this project. I am also grateful to the Institute for the Study of Human Rights at Columbia University for enabling me to construct this thesis.

Finally, I would like to thank Professor Tracey Holland for your supervision of this project. Your unwavering support and constructive feedback have been instrumental at every step, and your guidance has enabled me to achieve more in my research than I ever thought possible. Thank you.

I. Introduction

The Convention on the Rights of the Child obliges member states to protect children from all forms of abuse, including sexual exploitation (Art. 19) and trafficking (Art. 35). Article 2 also requires that states ensure the implementation of all rights put forth in the convention within their jurisdiction.[1] In addition to the convention, *the Children's Rights and Business Principles* were created to guide businesses' responsibilities to enact the articles of the convention. Although not legally binding, all corporations are encouraged to respect the principles and are obliged to comply with any related national laws.[2] In 2018, the United States passed the *FOSTA-SESTA* bill, which made internet companies in the U.S. liable for "knowingly" hosting sex-trafficking material on their platforms, including that of children.[3]

Despite this, the number of Cybertips the National Center for Missing and Exploited Children (NCMEC) received for online content containing sexually exploited children, including victims of trafficking, is on the rise. In 2015, the NCMEC reported receiving tips for 6.5 million pieces of sexually exploited content of minors, such as photographs and videos posted online. In 2017, that number rose to 20.6 million, and in 2019, they reported a staggering 69.2 million cases.[4] NCMEC also directly correlates the increased reports of sex-trafficked children to the growing usage of online advertisement as a method for pimps to locate potential customers.[5] A study found that 38% of victims trafficked before 2004 were advertised using online methods.

---

[1] United Nations, "The Convention on the Rights of the Child." *Treaty Series 1577* (1989): 1-15.
[2] United Nations International Children's Emergency Fund, "Children's Rights and Business Principles," (2012): 12-13.
[3] Elise Blegen. "Sex, Drugs, and Rock & Roll: Effectively and Equitably Moderating Vice and Illegal Content Online." *Vanderbilt Law Review* 76 no. 4 (2023): 1280.
[4] Blegen, 1269.
[5] Sandra Kowalski. "Holding Internet Advertising Providers Accountable For Sex Trafficking: Implements to Criminal Prosecution and a Proposed Response" *Public Interest Law Journal* 27 (2018): 101.

However, for persons victimized after 2004, this increased to 75%.[6] Because technology is becoming increasingly accessible and affordable, the problem is likely to get worse before it gets better.[7]

Due to the *Communications Decency Act* (1996), software companies have regulatory authority over the content published on their websites and are, therefore, in charge of filtering harmful posts from their domains.[8] Researchers have tested the content that viewers are likely to engage with and found that, although people tend to view content that aligns with their ideologies, there are circumstances where viewers are more likely to engage with posts that contradict their moral beliefs. This is a phenomenon referred to as the "confrontation effect," where highly charged topics become more engaging for the viewer,[9] as controversial content connects with the viewer's emotions and creates moral outrage.[10] As websites profit from selling advertisement spots[11] and keeping engagement high on their platforms to expose users to more ads,[12] rational behavior (profit maximization) will result in limited content moderation if external monitoring is not increased. In turn, monitoring requires increased transparency regarding the companies' content moderation practices.

---

[6] Vanessa Bouché. *Survivor Insights: the Role of Technology in Domestic Minor Sex Trafficking:* Thorn (2018).
[7] Eva Veldhuizen-Ochodnicanová. "Technology-facilitated Child Abuse, Exploitation and Trafficking" April 8, 2024, Johns Hopkins School of Advanced International Studies, Baltimore, Maryland, United States https://bipr.jhu.edu/events/4793-The-Human-Trafficking-Initiative-Series--Technology-facilitated-Child-Abuse-Expl.cfm.
[8] Blegen,1282.
[9] Daniel Mochon and Janet Schwartz, "The Confrontation Effect: When Users Engage More with Ideology-Inconsistent Content Online." *Organizational Behavior and Human Decision Processes* 185 (2024): 1.
[10] Mochon and Schwartz, 11.
[11] Mary Franks, "Justice Beyond Dispute," *Harvard Law Review* 131 (2018): 1374-1397.
[12] Mochon and Schwartz, 12.

In recent years, many tech firms have begun releasing transparency reports that reveal information about their internal content moderation guidelines.[13] In theory, this would increase the capability of external monitoring over content moderation. Yet, critics argue that these reports simply give the appearance of transparency instead of actually improving transparent behavior.[14] Transparency reports are essential for potential advertisers, who use them to understand the platform's approach to brand safety and security. This impacts decisions to purchase advertisement space or not on the website.[15]

My research centers on the transparency reports released by three companies that have been identified in the past to have hosted prostitution advertisements for sex-trafficked children: Meta (Facebook), MindGeek (Pornhub), and Snap Inc (Snapchat). I have compared their transparency reports to a useful human rights framework, the *Santa Clara Principles 2.0,* which human rights experts created in Santa Clara, California, in 2020 to guide the formatting of transparency reports, ensuring the protection of human rights.[16] I aim to identify areas in each company's transparent reporting that violate the requirements of the principles. By highlighting gaps and analyzing how improvements would affect child sex trafficking material online, I hope to contribute to the existing literature on ways to stop the increase of child sex trafficking material. To guide my research, I have developed the following research question: Why are

---

[13] Jennifer Cobbe et al., "Regulating Transparency?: Facebook, Twitter and the German Network Enforcement Act," Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, (New York: Association for Computing Machinery, 2020), 261-271.

[14] Mykola Makhortykh and Aleksandra Urman. "How Transparent are Transparency Reports? Comparative Analysis of Transparency Reporting Across Online Platforms," *Telecommunications Policy* 47 (2023), 1.

[15] "What is A Transparency Report," Trust and Safety Professional Association, accessed November 22, 2024. https://www.tspa.org/curriculum/ts-fundamentals/transparency-report/what-is-a-transparency-report/#:~:text=Transparency%20reporting%20can%20be%20important,for%20openness%20with%20the%20public.

[16] "The Santa Clara Principles On Transparency and Accountability in Content Moderation," Santa Clara Principles, accessed November 22, 2024, https://santaclaraprinciples.org/.

transparency reports insufficient as tools for monitoring content moderation of child sex trafficking material online?

II. Literature Review and Theoretical Frameworks

A. Children's Rights

The Convention on the Rights of the Child was adopted by the United Nations in November 1989 and entered into force in September 1990. It was developed based on the Universal Declaration of Human Rights, which recognised that children (referring to all persons under the age of eighteen) are entitled to special care. The convention emphasizes, through its 54 different articles, that children require special protection, security, and assistance to ensure that everyone has the right to grow up in a safe and nurturing environment, where they can develop the personality necessary for adulthood. [17]

However, the convention is not the only piece of international law that defines the importance of children's human rights. The International Covenant on Civil and Political Rights (Art. 23) emphasizes the significance of the family and requires the protection of the child in the event of dissolution of the family unit. Article 24 also proclaims that each child has the right to a name given to them at birth, a nationality, and special protections regardless of, for instance, race, sex, language, or social class.[18] General Comment 17 on Article 24 (Rights of the Child) further explains that the right to protection for the child shall be received from both the family, society, and the state, and that extends beyond the demands of Article 2, which requires the state to protect all human beings in its jurisdiction. For example, no child shall be subjected to capital punishment, and in the event of arrest, the child shall be held separately from adults. States should also take all necessary measures to reduce infant mortality rates, ensure access to education, and prevent malnutrition among children. In addition, the state is obliged to prevent

[17] United Nations, "The Convention on the Rights of the Child."
[18] United Nations (General Assembly). "International Covenant on Civil and Political Rights," *Treaty Series* 999 (1966): 171.

violence against children, such as preventing their exploitation in warfare or as labor power. Finally, the General Comment stipulates that children shall be awarded all of these protections due to their status as minors.[19] In other words, it is precisely because they are children that they deserve specifically defined rights.

In 1946, the United Nations International Children's Emergency Fund (UNICEF) was established to protect the rights of children worldwide. The organization originally emerged in response to the unprecedented refugee crisis that followed the end of the Second World War. UNICEF was created by the United Nations Relief Rehabilitation Administration to help children in the aftermath of the war. In the 1950s, the organization's mandate was expanded to assist children outside European borders, with a focus on programs for clean water, sanitation, and nutrition. In the 1960s, the mission was expanded to include a focus on every child's right to education, and in the 1970s, they began emphasizing children's right to healthcare.[20] During the 1980s, UNICEF launched missions promoting immunisation programs and oral rehydration therapy.[21] At the turn of the 21st century, the organization convened the World Summit for Children and launched extensive campaigns to combat AIDS.[22] To this day, the organization remains a leading voice in the fight for children's rights worldwide and has, for example, helped distribute vaccines following the COVID-19 pandemic.[23]

UNICEF further explains why children's rights are significant to distinguish from human rights. Children are not part of their parents, communities, or states. Instead, they are individuals,

---

[19] United Nations Human Rights Committee (HRC), CCPR General Comment No. 17: Article 24 (Rights of the Child), (1989): 1-3.
[20] United Nations International Children's Emergency Fund, "Emerging from the ashes of war: 1946–1979," accessed April 17, 2025. https://www.unicef.org/stories/learning-experience-19461979.
[21] United Nations International Children's Emergency Fund, "Moving with the times: 1980–1988," accessed April 17, 2025. https://www.unicef.org/stories/learning-experience-19801988.
[22] United Nations International Children's Emergency Fund, "Bringing the world together: 1989–2005," accessed April 17, 2025. https://www.unicef.org/stories/learning-experience-19892005.
[23] United Nations International Children's Emergency Fund, "Becoming the leading voice: 2006–2020," accessed April 17, 2025. https://www.unicef.org/stories/learning-experience-20062015.

not possessions, and as such, they have equal status in terms of human dignity as adults. Children are also dependent on adults during the developmental years, and therefore, the state has additional obligations to care for children whose family units or caregivers cannot meet the necessary standards and the needs of the child. This is even more significant because children who are still developing and growing are more sensitive than most adults to poor living conditions, including inadequate healthcare and malnutrition. In addition, most areas of society, such as public health, affect children, but they often do not participate in the political environment and are unable to vote. Therefore, additional attention must be spent on children to ensure that their opinions and needs are heard. Finally, the organization emphasizes the research, which shows that societies that neglect the needs of their children suffer grave consequences as the experiences early on in life determine who the child will become and their cost or contribution to the future development of society.[24]

B. Sex Trafficking of Children and the Internet

Trafficking in all forms is a violation of human rights. International law prohibits "practices associated with human trafficking" such as debt bondage, forced labor, enforced prostitution, sexual exploitation of children, and slavery.[25] What constitutes human trafficking is defined by the United Nations *Protocol to Prevent, Suppress and Punish Trafficking in Persons, Especially Women and Children* (Palermo Protocol). It puts forth the requirement for the presence of an *act*, *means*, and *purpose* to consider a case to be human trafficking. As such, trafficking is understood as an *act* of ". . . recruitment, transfer, harboring, or receipt of persons. .

[24] United Nations International Children's Emergency Fund, "Child Rights and Why They Matter: Every Right, For Every Child," accessed April 17, 2025. https://www.unicef.org/child-rights-convention/child-rights-why-they-matter.
[25] United Nations Human Rights Office of the High Commissioner. "Human Rights and Human Trafficking" *Fact Sheet No. 36* (2014): 5.

.," by *means* including force, coercion, or abduction for the *purpose* of exploitation.[26] The protocol established that when any of the listed *means* are present, the consent of the victim shall be irrelevant. Consequently, during the prosecution of a trafficker, the defendant may not argue that their victims are consensual sex workers. On the other hand, in cases of trafficking in children (persons under the age of eighteen), consent cannot be legally given. As such, there is no requirement for the presence of *means* for the case to be considered human trafficking. Instead, it is defined as an *act* with the *purpose* of exploiting the child.[27] Therefore, anyone who financially gains from or pays for the sexual exploitation of a child online can be argued to have engaged in the trafficking of that child. This includes selling or engaging with child pornography.[28]

The Convention on the Rights of the Child also establishes legally binding protection of children from trafficking. As mentioned, Article 35 protects children from all forms of sexual exploitation. This includes the sexual coercion of a child, prostitution, and all types of child pornography and sexually exploitative material. Article 11 also obliges states to protect children from unlawful transfer abroad, and Article 32 requires that states prevent the economic exploitation of children.[29] Another legally binding source of international law prohibiting trafficking is the *Convention on the Elimination of All Forms of Discrimination against Women*

---

[26] United Nations, *Protocol to Prevent, Suppress and Punish Trafficking in Persons, Especially Women and Children, supplementing the United Nations Convention against Transnational Organized Crime*, 2237 U.N.T.S. 319 (2000).
[27] Beatriz Uittis. *Sex Trafficking of Children Online: Modern Slavery in Cyberspace* (Lanham: Rowman & Littlefield, 2022), 67.
[28] "How Facebook and Instagram Became Market Places for Child Sex Trafficking," *Guardian,* Apr. 27, 2023. https://www.theguardian.com/news/2023/apr/27/how-facebook-and-instagram-became-marketplaces-for-child-sex-trafficking.
[29] United Nations, "Convention on the Rights of the Child."

(1979), where Article 6 demands that states take all appropriate measures to prohibit trafficking of women and forced prostitution.[30]

The process of trafficking in children can be divided into two subparts: (1) the recruitment and (2) the distribution (advertising and selling) of a child. Both these aspects have changed drastically over time due to the introduction of the Internet.[31] In 1988, the "Meat Rack Report" (a five-year study of the different markets for child sex trafficking) mapped out the structures of sex trafficking in children at the time. Recruitment tactics were found to often span over multiple weeks. In the first stage, some pimps engaged in a seduction process of the victim where they showered the child with gifts, compliments, and attention. In the second stage, the pimp began bringing the child to parties and introduced them to the idea of having sex with adults by exposing them pornographic images and videos.[32] Other pimps lured victims by offering security,[33] paying off debts,[34] and entering relationships with the minor.[35]

Almost four decades later, the recruitment process for victims of child sex trafficking has dramatically changed. Perpetrators still often contact their victims directly and engage in one-on-one dialogue with the child. However, the Internet enables perpetrators to contact multiple children simultaneously and adopt different personas with each child, depending on what they identify that child to be receptive to.[36] The perpetrator is also protected from possible identification through encryption technology such as VPN-based services, cryptocurrency

---

[30] United Nations, "Convention on the Elimination of All Forms of Discrimination Against Women," *Treaty Series 1249* (1988).
[31] Veldhuizen-Ochodnicanová, Presentation.
[32] Daniel Campagna and Donald Proffenberger. *The sexual trafficking in children: an investigation of the child sex trade. (*Dover, MA, Auburn House, 1988), 97.
[33] Campagna and Proffenberger, 98.
[34] Campagna and Proffenberger, 99.
[35] Campagna and Proffenberger, 91.
[36] Nuria Lorenzo-Dus. *Digital Grooming: Discourses of Manipulation and Cyber-Crime,* (Oxford: Oxford University Press, 2022). 60-61.

transactions, and networks like Tor.[37] Consequently, even though the number of adults interested in children remains stagnant (around three to five percent of the population), the number of children they can target has exponentially increased with the introduction of the World Wide Web. Technological advancements in artificial intelligence have also created a new arsenal for perpetrators who, for instance, use AI-generated pictures to pressure children to send real images and videos of themselves in a process commonly referred to as "sextortion."[38] These files are later distributed amongst networks of exploiters, and the child is referred to other interested parties.[39] Before this transformative technology, the period between initial contact with a potential victim and exploitation of that child could entail months of grooming, persuasion, and coercion. Today, due to technological advancements, this period can be as short as nineteen seconds.[40]

Similarly, there has been a significant shift in the mechanisms for distributing advertisements for trafficked children. Before the rise of the internet, typical methods used by pimps to locate customers included handing out child pornography to potentially interested parties,[41] using personal connections and reputation,[42] hiring middle hands, such as bartenders, to refer potential buyers,[43] and placing the victim directly on the streets.[44] Today, the internet has created new pathways for abusing, exploiting, and trafficking children.[45] Child Sexual Abuse Material (CSAM), including ads for sex-trafficked children, is distributed through live streaming

---

[37] Brian Levine. "Increasing the Efficiency of Investigations of Online Child Sexual Exploitation." *University of Massachusetts Amherst*, (2022), viii.
[38] Veldhuizen-Ochodnicanová, Presentation.
[39] Levine, 40.
[40] Veldhuizen-Ochodnicanová, Presentation.
[41] Campagna and Proffenberger, 118.
[42] Campagna and Proffenberger, 106.
[43] Campagna and Proffenberger, 58.
[44] Campagnaand Proffenberger, 61.
[45] Veldhuizen-Ochodnicanová, Presentation.

services, internet forums, gaming platforms, and peer-to-peer file sharing.[46] This possibility of advertising and selling the child online also increases the child's isolation and the pimp's capability of remaining in the shadows.[47] With the rise of social media platforms, the methods for advertising a victim and connecting the child with a buyer have further changed. Platforms such as Facebook and Instagram alow pimps to locate potential buyers through publishing prostitution advertisements for sex-trafficked children. On these websites, it is possible to create a firm relationship between perpetrators and victims through direct messaging functions, which also allows pimps to directly discuss prices and locations with buyers.[48]

C. Self-regulation and Content Moderation

In 1996, the *Communications Decency Ac*t was passed by Congress in an attempt to protect free speech online. Section 230 of the act protects websites posting independent third-party content from liability for what is shared on their platforms. It also encourages companies to self-regulate harmful material as it prevents liability for partial editing or filtering the content. Before Section 230, websites could be held liable for attempting to moderate some of the content they hosted, as editing made the website liable for all material on their platform. However, as websites that engaged in no editing remained protected, a boycotting epidemic of moderation was ensured. As such, Section 230 was a vital legal progression in creating the internet we are today accustomed to.[49]

However, the immunity clause of Section 230 ultimately ended up protecting websites that hosted advertisements and other material linked to child sex trafficking, as full regulatory

[46] Levine, 12-13.
[47] Kowalski, 100.
[48] "How Facebook and Instagram Became Market Places for Child Sex Trafficking."
[49] Blegen, 1275.

power was given to the companies themselves.[50] The *FOSTA-SESTA* bill (2018) was enacted to combat this development by amending the immunity clause. It establishes that immunity will not be granted to platforms that "knowingly" assist, facilitate, or support sex trafficking material, including that of children, on their platforms. However, to this date, the *FOSTA-SESTA*, with its high promises to combat sex trafficking online, has never been utilized in a prosecution.[51] Hence, despite these legal changes, the problem of sex trafficking of children online continues to increase.[52] Experts in the field have also identified one underlying cause of the troubling increase to be the continued lack of self-regulation amongst websites hosting child sex trafficking material.[53]

D. The Principal-Agent Theory and Content Moderation

To understand why companies continue to limit their content moderation reporting, we can apply Economic Theory. The Principal-Agent Theory, which Michael Jensen and William Meckling put forth, explains why corruption occurs in the market. The theory identifies *Principals* as actors who outsource jobs to *Agents*. The *Principal* wants the outsourced task to be completed to the best possible outcome for the lowest price, while the *Agent* wants the least amount of work for the highest pay. Therefore, a conflict of interest occurs, and the theory states that since the cost of enforcement and monitoring is high for the *Principal*, the *Agent* will engage in "shirking." This is opportunistic behavior, such as engaging in embezzlement, fraud, and other corrupt practices in the pursuit of profit maximization.[54] In the case of limiting prostitution advertisements online of sex-trafficked children, the task of moderating the ads has been

[50] Blegen,1282.
[51] Blegen, 1280.
[52] Uittis, 97.
[53] Veldhuizen-Ochodnicanová, Presentation.
[54] Claudia Halbac. (Lecture, American Economy at Columbia University, New York, NY, October 15, 2024).

outsourced by the *Principal* (the U.S. Government) to the *Agent* (the software company). However, due to a lack of enforcement mechanisms and high monitoring costs, the software company is likely to engage in shirking to increase its profits. In turn, many online platforms do not profit from selling products but from selling advertisement spots. As such, the existence of advertisements for victims of sex trafficking might not be a "bug in the system; rather a feature."[55] In other words, overly strict content moderation of advertisements is undesirable for the company.

Similarly, excessive transparency regarding content moderation is also undesirable for companies due to a phenomenon known as the "Streisand Effect." This refers to the fact that being overly transparent about removing a piece of harmful content (like CSTM) draws attention to the content in the first place. This can create a climate where the presence of CSTM appears even more extensive than it is.[56] As discussed previously, potential advertisers refer to companies' transparency reports to get a better understanding of a platform's climate before deciding to purchase advertisement spots.[57] If the report includes extensive transparency about CSTM, the advertiser may fall victim to the "Streisand Effect" and, therefore, decide not to advertise on the platform. Consequently, the Agent (the software company) is further incentivised not to exhibit greater transparency regarding their content moderation process.

The Principal-Agent theory has been mirrored in investigations into websites known to host CSTM and engage in content moderation. For instance, Backpage.com, a website that hosted escort advertisements under its "Adult Section," was connected to 73% of all claims of child sex trafficking NCMEC received from the general public in 2017. That same year, the website reportedly made 99% of its global revenue from advertisements posted under the "Adult

[55] Franks, 1374-1397.
[56] James Grimmelman. "The Virtues of Moderation," *Yale Journal of Law & Technology* 42 (2015): 66.
[57] "What is A Transparency Report."

Section."[58] In 2015, it was reported that Facebook was home to the second most ads for sex-trafficked children after Backpage.com,[59] and in 2019, the site removed over 11 million pieces of sexually exploitative content of children in less than three months.[60] Another site known to have hosted material of sex trafficked children is Pornhub which in 2020 had over 3.5 billion monthly visitors, which is more than Netflix had at the time.[61] Shortly after a New York Times article came out revealing the hub of child pornography Pornhub was, the company removed over eight million videos overnight, representing 65% of the website's content.[62] This indicates an awareness of what uploaded videos potentially contained rape of children.

E. Transparency: A Method of Combatting Corruption

An essential part of anti-corruption research is understanding the role of transparency as a balancing mechanism to corruption. Transparency can help reduce corruption by decreasing the resources required for corrupt behavior. This is achieved by making information available to the public, thereby removing the information monopoly as a tool used by corrupt actors to extort, embezzle, or cover up their actions and behaviors. Transparency, and as such, increased access to an *Agent*'s internal information, also lowers the costs of external and third-party monitoring, thereby decreasing the risk of the *Agent* engaging in corruptive behavior. [63] In the case of content moderation practices, we are predominantly looking at corporate transparency, which refers to private actors' willingness to voluntarily provide information about internal data to their

[58]Jacqueline Hackler. "Inconsistencies in Combatting the Sex Trafficking of Minors: Backpage's Deceptive Business Practices Should Not Be Immune from State Law Claims." *Seattle University law Review* 40, (2017), 1109.
[59] Vanessa Bouché. *A Report on the Use of Technology to Recruit, Groom, and Sell Domestic Minor Sex Trafficking Victims:* Thorn (2015). 19.
[60] Blegen, 1291.
[61] Nicholas Kristof. "The Children of Pornhub." *New York Times*, Dec. 4, 2020.
[62] Blegen, 1269.
[63] Mungiu-Pippidi, 3-4.

stakeholders or share it under obligation by a state.[64] Many studies point to the advantages of increasing the transparency of a company and ensuring that stakeholders have greater access to internal data, as it decreases information asymmetry.[65] Access to this knowledge also allows members of civil society to hold the agent accountable for behavior. Simultaneously, it enables researchers and legislators to access the required information for informed decision-making and to produce quality research.[66] Therefore, it is not strange that multiple researchers have pointed to the problematic nature of the minimal transparency reporting among companies regarding their exact content moderation practices.[67]

Many social media platforms and software companies today release transparency reports, allowing outsiders to examine the inner workings of the company's content moderation practices and increase accountability for any corrupt behavior. The reports contain voluntarily published information about the material that the company has moderated. This can include data about the number of posts that have been removed, guiding principles, and any other information the company wishes to disclose. In theory, creating these reports shows a step towards increased transparency. However, how "transparent" these transparency reports actually are is up for debate.[68] Some researchers have argued that these reports allow companies to display a false illusion of transparency. Others have found that the reports attract advertisements to large software companies' platforms as potential advertisers study the reports to understand company approach to brand safety and to identify what type of user environment the platform hosts[69] In

---

[64] Etienne Farvaque and Catherine Refait-Alexandreand and Dhafer Saïdane. "Corporate Disclosure: A Review of its (Direct and Indirect) Benefits and Costs," *International Economics* 128, no. 4. (2011): 5-32.
[65] Ardigó, 2.
[66] Mungiu-Pippidi, 4.
[67] Paul Rosenzweig and Katie Stoughton. "Toward Greater Content Moderation Transparency Reporting." *Lawfare (*2022).
[68] Makhortykh and Urman, 1.
[69] "What is A Transparency Report," Trust and Safety Professional Association, accessed November 22, 2024.

other words, scholars have argued that the transparency reports do not guarantee that the company is implementing transparent practices. Instead, their role is to create a beneficial narrative for the company.[70]

F. Automatic Content Moderation and Fighting CSTM Online

Automatic content moderation tools are becoming increasingly utilized in identifying and moderating child sexual abuse material online. They enable companies to moderate large numbers of content and accounts quickly.[71] One of the most commonly applied automatic detection and moderation systems in the fight to prevent the spread of child sexual abuse material, including child sex trafficking material, is *hash-sharing*. This method identifies an image's unique digital fingerprint, which is referred to as the image's hash. Algorithms are implemented that actively search for hashes belonging to previously identified CSAM material for companies to be able to stop the spread of the content and quickly remove it from their platforms.[72] The National Center for Missing and Exploited Children (NCMEC) has gathered an extensive hash-sharing database where companies can submit identified CSAM material from their platforms, which other internet companies can then access and program their own algorithms to identify.[73] As such, CSAM, if located on one platform, can be prevented from spreading to other websites.

Human moderators can identify harmful material and take action against it, but, as explained above, software can moderate at a far greater speed and volume. However, some

---

https://www.tspa.org/curriculum/ts-fundamentals/transparency-report/what-is-a-transparency-report/#:~:text=Transparency%20reporting%20can%20be%20important,for%20openness%20with%20the%20public.
[70] Cobbe et al., 261-264.
[71] Clare Cho and Ling Zhu. "Social Media: Content Dissemination and Moderation Practices," *Congress.gov* (2025): 1.
[72] Rachel Haney. "When 'Safe at home' is Not Safe," *Scitech Lawyer* 17, no. 4. (2021): 22-27.
[73] "CyberTipline 2023 Report," National Center for Missing and Exploited Children, accessed April 17, 2025, https://www.missingkids.org/cybertiplinedata.

researchers have raised concerns about using automatic content moderation in the context of, for example, pornography and sexually exploitative material of minors.[74] Automatic moderation systems are more rule-bound than humans, which makes them a suitable tool for moderating content that clearly violates programmed guidelines. It is also more cost-effective. However, it does not understand nuances to the same extent as people, and as such, a company will end up having to make a trade-off between quality and cost when deciding to implement automatic decision-making software instead of using human moderators.[75] Governments are also raising concerns that these systems are flawed and may not remove content in accordance with state laws. Due to the continued lack of transparency regarding how and when these automatic detection tools are used, further reliance on automatic content moderation tools may worsen the problems rather than improve them.[76]

[74] Grimmelman, 64.
[75] Grimmelman, 65.
[76] Cho and Zhu, 13

III. Methodology

A lack of content moderation continues to be a significant issue in preventing sexual exploitation of children online.[77] To hold companies accountable for the actions they take against the spread of child sex trafficking material (CSTM), transparency about their moderation systems is necessary. However, some researchers have argued that the reports companies publish to prove transparent moderation behavior display a false illusion of openness.[78] Therefore, my research aims to understand why transparency reports are insufficient as tools for monitoring content moderation of child sex trafficking material online.

After engaging with the existing literature and looking at two reports of child sexual abuse material and child sex trafficking advertisements produced by Thorn and NCMEC, I investigated three companies where researchers have identified the presence of CSTM in the past: Meta (Facebook), MindGeek (Pornhub), and Snap Inc. (Snapchat). By studying the companies' transparency reports, I hope to identify where each company is not complying with international standards, specifically the *Santa Clara Principles on Transparency and Accountability Around Content Moderation 2.0.* Increasing transparency regarding their content moderation practices will enable the public and law enforcement to get access to the information necessary to hold companies accountable for their commitments to combat the spread of CSTM on their platforms.[79]

A. Where to Locate the Transparency Reports

Because there are no universal standards for transparency reporting, all companies decide what information they wish to disclose in their published reports, as well as how they format and

[77] Veldhuizen-Ochodnicanová, Presentation.
[78] "What is A Transparency Report."
[79] Mungiu-Pippidi, 4.

structure them.[80] Pornhub presents its transparency reports under its "Trust and Safety" section of help.pornhub.com. Here, you can access annual reports from 2021 and biannual transparency reports starting in 2023. Pornhub publishes one (maximum two) report per year that contains all different areas of the content moderation process for which they report data. You can also read Pornhub's mission statement for transparency reporting in their first report on the website for 2020, which lays out the guiding principles they follow, their intention, and some early data regarding their content moderation systems.[81]

Similarly to Pornhub, Snapchat shares its transparency reports at https://values.snap.com, where they list biannual reports dating back to 2014. Until 2021, each publication consisted of a single, comprehensive report that summarized their content moderation efforts on a global level, as well as presenting some country-specific data. However, after 2021, Snapchat began releasing country-specific reports for 53 nations alongside one shorter global report.[82]

Finally, the website transparency.meta.com/reports/ compiles all of Facebook's published transparency reports. Contrasting Pornhub's and Snapchat's transparency reporting structure (releasing one report annually or biannually), Facebook publishes reports under seven different thematic categories.[83] Multiple shorter reports are shared under these seven different areas, all following different publishing schedules. Some reports are released quarterly, while others are distributed monthly, biannually, or annually. Therefore, to understand Facebook's transparency reporting within set time intervals, one must read through multiple different reports covering different themes of content moderation during the same period.

[80] Makhortykh and Urman, 2.
[81] See Appendix F for a list of all reports they have published so far.
[82] See Appendix F for a list of all reports they have published so far.
[83] See Appendix G for detailed information.

B. The *Santa Clara Principles 2.0* and its Role in Evaluating Transparency Reporting

The *Santa Clara Principles on Transparency and Accountability Around Content Moderation 1.0* were created in 2018 by several human rights organizations, scholars, and experts, as concerns for a lack of transparency amongst software companies increased. These principles served as guidelines for what internal information should be published to allow the public to get a greater understanding of internal content moderation practices. Many large companies, including Meta and Snap Inc., endorsed these three original principles. However, as the role of transnational internet corporations continued to develop, the original three principles were extended. In 2020-2021, the *Santa Clara Principles 2.0* (SCP 2) was created, which detailed what companies should do to ensure maximum transparency in the changing climate. SCP 2 is split into two parts: 1) Foundational Principles, which all companies should comply with, and 2) Operational Principles, which only large and well-established companies are expected to implement. The list contains five Foundational Principles, three Operational Principles, and two principles targeting state responsibilities.[84]

The principle which applies to my research into Pornhub's, Snapchat's, and Facebook's transparency reports is Operational Principle One, *Numbers*.[85] This principle contains 23 different requirements for data that each company ought to include in their reports to ensure transparency.[86] I have chosen not to use the Foundational Principles as they are guidelines for transparent behavior outside of content moderation transparency reports. This includes presenting user guidelines in a comprehensible way, ensuring that moderators are aware of

[84] "The Santa Clara Principles On Transparency and Accountability in Content Moderation."
[85] I will not be looking at the two principles targeting state governments nor the two last principles listed under the Operational Principles: *Notice* and *Appeal*. These two principles are implemented only after a piece of content has been moderated, including the user's interaction with the company. Therefore, these two principles are not relevant to studying a transparency report for the purpose of examining CSTM.
[86] See Appendix H for a detailed list of each of the 23 points.

cultural differences, and mitigating discrimination in the content moderation process.[87] The data points mentioned in the Foundational Principles that should be included in transparency reports are reiterated in the 23 requirements of Operational Principle One, *Numbers.* Hence, by relying solely on this principle, I am also able to cover the aspects of the Foundational Principles.

C. Creating a Table and Point System for Analysing and Collecting Data

My goal was to compare all published transparency reports from Facebook, Pornhub, and Snapchat with the *Santa Clara Principles 2.0*. Therefore, I needed to create a table and a point system to compare each company to the SCP 2.0, how they have changed over time, and how they compare to each other. I decided to rely on a simple point system to generate data that was easy to compare and analyse. If a report fully complies with one of the requirements listed in the *Numbers* principle,[88] I award the report 1 point. If it partly complies, it receives 0.5 points.[89] Finally, if the report does not include the required data points, it receives 0 points.

Following the decision to use this straightforward point system, I needed to construct a format for presenting my data. Firstly, I compared each requirement of the *Numbers* principle with the reports published by all three companies. Therefore, I created three separate tables, one for each company, which are presented in Appendices A to C. Each Table has 23 rows, and each one represents one of the 23 requirements listed in the *Numbers* principle.[90] Each column on the tables represents one separate report published by the company, where number one is the oldest report, number two is the second oldest, and so on.[91] Secondly, to visualize the findings from my three tables, I created three separate stack charts (Fig. 1, Fig. 2, and Fig. 3). Along the y-axis of

---

[87] "The Santa Clara Principles On Transparency and Accountability in Content Moderation."
[88] See Appendix H for a detailed list of each of the 23 points.
[89] See Appendix G for an explanation for why a report only received 0.5 points.
[90] See Appendix H for a detailed list of each of the 23 points.
[91] See Appendix A, B, and C for the tables and Apeendix E for a detailed list of each published report by each company.

each chart are the total scores a company has received for all of its published transparency reports to date. The x-axis shows the 23 requirements of the *Numbers* principle. There is also a secondary axis title for the x-axis showing the division of the 23 requirements of the *Numbers* principle into four categories: a), b), c), and d). These represent the four categories that the 23 requirements are officially divided into under the *Numbers* principle: a) *general reporting,* b) *state involvement*, c) *flagging system*, and d) *automated decision-making tools*.[92] The charts further distinguish these four categories (Fig. 1, Fig. 2, and Fig. 3) by three dotted black vertical lines.

There are two other important things to mention about the tables. Firstly, as previously discussed, Pornhub (MindGeek) began releasing transparency reports in 2021, and up until 2023, they only released one report per year. However, in 2023, they began publishing biannual reports, so two reports are available for 2023. Then, in 2024, they released the EU DSA Transparency Report alongside their regular reports. As such, they published a total of four reports during 2024. To conduct a comparative analysis of the three companies' content moderation transparency reporting, I had to use the same framework for each company to avoid skewing the data. Snapchat publishes its transparency reports biannually, meaning it publishes one report for the first half of the year and another for the second half. This matches Pornhub's reporting structure only during 2023, where they published two biannual reports. To ensure I follow the same structure as Snapchat, I have taken the two annual reports they published (for 2021 and 2022) and counted each report twice to create two biannual reports for each year. Therefore, when examining the tables in Appendix B, columns one and two have identical scores, as they are based on the same annual report. The same applies to columns three and four.

[92] See Appendix H for a detailed explanation of why the SCP has divided all the requirements under the *Numbers* principle into four categories.

Additionally, in 2024 Pornhub released four separate reports; two for the first six months of the year and two for the last. To follow the same biannual report formula, I have simply combined the annual reports covering the same six months. For example, during the first six months of 2024, Pornhub released two separate reports (the EU DSA Transparency Report and their regular biannual report). For requirement 1 out of the 23 requirements under the *Numbers* principle, the EU DSA report scored a 0.5, and the regular report received 1 point. Therefore, I have awarded the cumulative biannual report, which encompasses the information from both reports, a score of 1.

Secondly, as previously mentioned, Meta (Facebook) does not publish one consecutive report covering all areas of content moderation for which they provide data. Instead, they divide it into thematic areas where reports are released following different time schedules. Consequently, I had to create my own time parametric for this study to be able to compare Facebook's reports to those of Snapchat and Pornhub. I decided to develop biannual time cutoffs ranging from their first published report in 2013 to their last report in 2024. Each biannually published report neatly falls into one of these constructed reports. However, information published in an annual report is applied to both biannually constructed reports for this study. On the other hand, a report published monthly or quarterly falls within one of the constructed reports and is, as such, applied to the appropriate biannually constructed report. Therefore, each report I built for my table[93] contains information pulled from multiple published reports that all fall within the same six months.

Finally, the companies have all continued to increase the information about their content moderation practices that they publish since their initial reports. As such, I wanted to investigate if the fact that they are releasing more information means that they are becoming more

[93] See Appendix C for detailed information.

transparent in their content moderation processes. Therefore, I wanted to determine if the increase in published information also corresponded to an increase in the content moderation score in this study. If the score rises proportionally to the amount of information they publish, I can conclude that they are attempting to become more transparent. However, if I found that they have published exponentially more information but without their transparency score increasing, then I can conclude that an increase in information does not necessarily represent an increase in transparent behavior. Instead, it may indicate that the companies are releasing more information to appear more transparent, without publishing the information that the *Santa Clara Principles 2.0* have deemed necessary to ensure transparent behavior among large corporations.

Consequently, I created three graphs (Fig. 4, Fig. 5, and Fig. 6) showing the relationship over time between the amount of information published by each company and their transparency score. The y-axis of each graph shows the percent increase since year zero (the year they published their first report). Along the x-axis, each biannual report published since year zero is listed. Each graph has a line representing the percentage increase in the total transparency score the company received each year, which I calculated by adding the scores it received for each of the 23 requirements every six months. Another line on each graph shows the percent increase in published information since year zero (the blue line). In the case of Snapchat and Pornhub, I have found this statistic by counting the number of words included in each report. For Facebook, which has released a distinctly separate number of reports each year, I have instead counted the number of published reports for each biannual time parametric.

D. Limitations

There are multiple limitations to my study. Firstly, although based on international standards, the point system will ultimately render a subjective evaluation outcome. However, I

have tried to mitigate this by only awarding half points when the data is not fully included, rather than evaluating the extent to which it is reported and rewarding, for example, 0.3 points or 0.7 points. A point system like that requires conducting a value-based analysis of the data, where I compare the extent to which I believe the data in one report was represented compared to the other. By creating a more straightforward system, some nuances are lost. However, I can remove some of my biases from the project.

Secondly, in the case of Meta (Facebook), I have only considered the reports listed in the previous section of this paper. Some of the areas of thematic reporting contain links for additional reports, which in turn contain links for even more reports, and so on. As such, I may have missed data points reported in further reports. Additionally, some of these reports are not available in English, such as the *Brazil TSE Election Rules Transparency Reports* (August 2024 to December 2024). Finally, the world of content moderation is rapidly changing, meaning that the subsequent updates made to each company's transparency website or the following report itself may look significantly different, and the gaps I have been able to identify may be filled. In other areas, backsliding might have occurred. Therefore, the findings of this study should be looked upon in the context of time (my data was collected in December 2024 and January 2025) and understood to represent the current climate of transparency reporting and prostitution advertisements for sex-trafficked children.

IV. Findings and Analysis

The three companies in this study, which have all pledged to combat child sex trafficking material on their websites, implement automatic content moderation tools extensively. These tools enable companies to moderate a large amount of content and accounts quickly,[94] and allow for the instant detection of child sex trafficking material (CSTM) that has already been identified on other websites using digital fingerprinting methods.[95] This is a significant aspect of their content moderation processes. As such, if companies wish to be transparent about these processes, information about automatic content moderation tools would be included in their transparency reports. Yet, this study finds that not to be true.

A) Snapchat Refrains from Including Information about Automated Moderation Tools

When comparing Snapchat's published transparency reports to the *Santa Clara Principles 2.0* Operational Principle One, *Numbers,* I found that Snapchat continues to score the lowest in category d) *automated decision-making tools.* Figure 1 depicts how many points (shown on the y-axis) Snapchat has scored across each of the 23 requirements listed by the *Numbers* Principle[96] (shown on the x-axis). Snapchat can receive a maximum of 1 point per published report for each requirement (1 to 23). Since they have published twenty reports, they can receive a maximum of 20 points per requirement (1 to 23 on the x-axis). Hence, a score of 20 points indicates that they included the requirement in each report they have published. A score of 10 points means that they included the requirement in half of their published report, and so on. Moreover, the *Santa Clara Principles 2.0* has divided the 23 requirements listed in the *Numbers* Principle into four categories: a) *general reporting,* b) *state involvement,* c) *flagging system*, and d) *automated*

---

[94] Cho and Zhu, 1.
[95] Haney. "When 'Safe at home' is Not Safe."
[96] See Appendix H for a detailed list of each of the points in the *Numbers* principle.

decision-making tools. Therefore, Figure 1 displays a secondary x-axis title below the 23 requirements, which lists the four different categories. To further distinguish each category in the graph, I have included three black dotted lines separating the requirements into their respective groups.

Figure 1 reveals that Snapchat scored the lowest in category d) *automated decision-making tools.* The company received 0 points for requirements 19 to 22, which ask for the inclusion of the accuracy rates of automatic moderation tools, the process of appeal, the criteria for when they are implemented, and the extent of human oversight over the automatic processes. Snapchat only received 0.5 points for requirement 17 and 1 point for requirement 18, which asks when and how these systems are implemented and for what specific categories they are used. Only for requirement 23 did Snapchat receive higher points (9 out of a possible 20), which means they included the information in less than half of their published reports. This requirement is the only one in category d) *automated decision-making tools,* which does not specify the need to include data. Instead, the company must only mention the different cross-platform automatic detection tool programs in which they participate, such as NCMEC's Hash Sharing Base. As such, it is the only requirement that does not ask for the inclusion of specific data, and it is the only requirement that Snapchat included in more than one report. Consequently, Snapchat continues to refrain from transparency regarding the implementation of AI and automated systems in its content moderation practices.

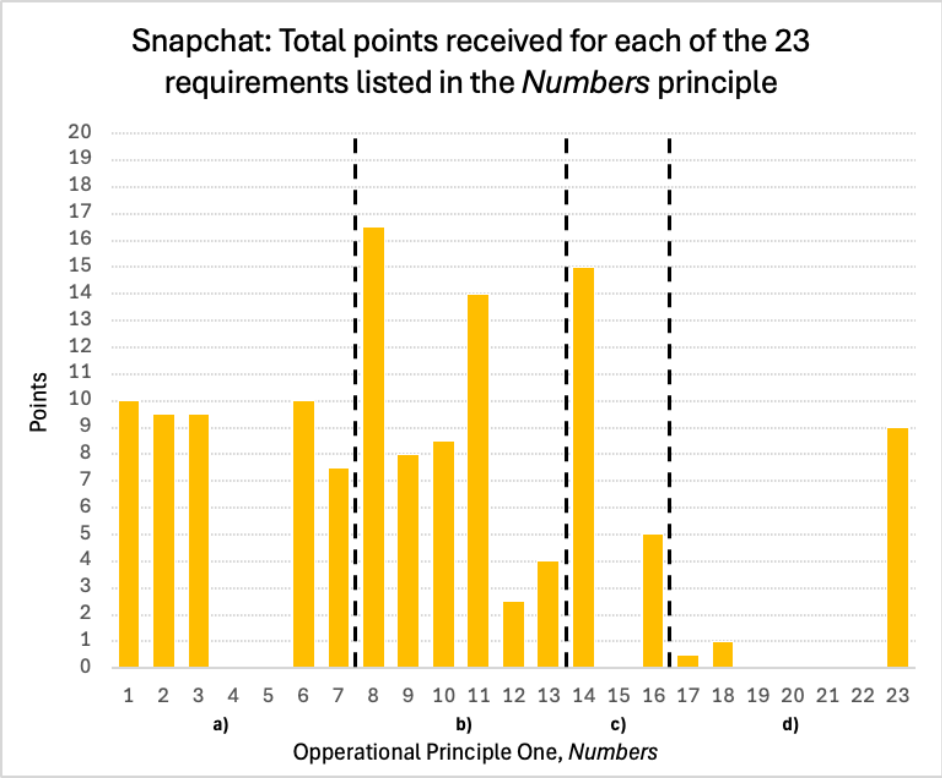

Fig. 1: Comparing all of Snapchat's reports with the 23 requirements in the *Numbers* Principle

B) Facebook Refrains from Including Information about Automatic Moderation Tools

Similarly to Snapchat, Facebook performed the worst in category 4) *automated decision-making tools,* as seen in Figure 2. This chart is constructed using the same logic as Figure 1. However, since Facebook has published four more reports than Snapchat, the maximum number of points (shown on the y-axis) Facebook can receive for each of the requirements listed along the x-axis is 24. Their highest score in category d) *automated decision-making tool* was 3.5 out of a possible 24 points for requirement 20, which asks for the automated processes' confidence, accuracy, and success rates. In other words, Facebook largely excluded information such as when and how they implement automatic moderation tools, the extent of human oversight, whether appeals were successful, and the types of cross-platform programs in which they participate.

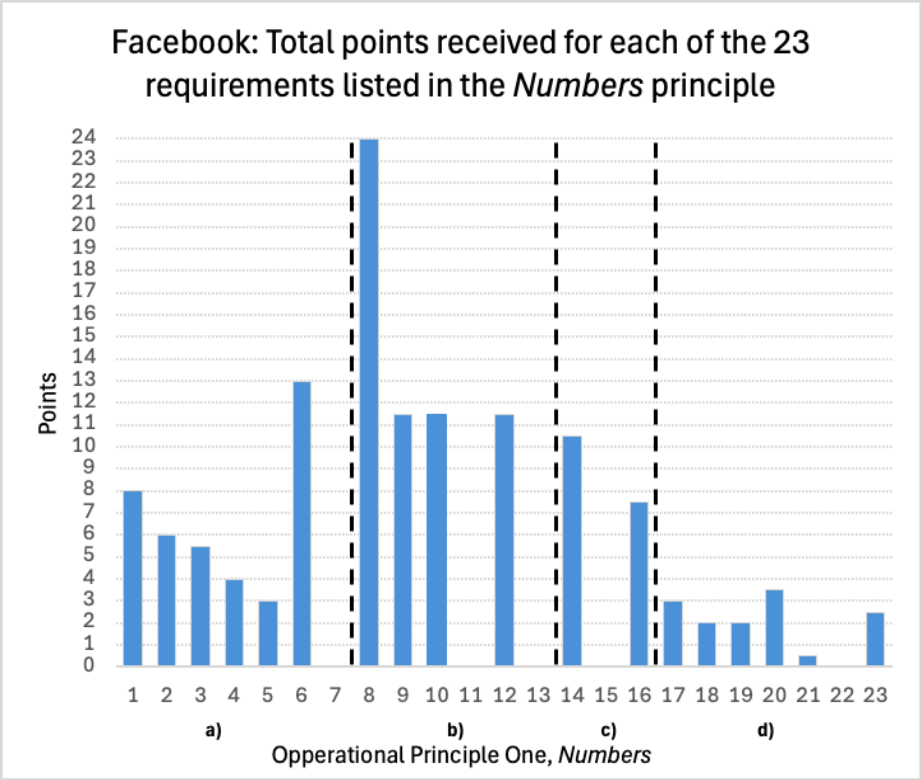

Fig. 2: Comparing all of Facebook's reports with the 23 requirements in the *Numbers* Principle

C) Additional Proof of Facebook's and Snapchat's Exclusion of Data Regarding Automatic Content Moderation Processes

By comparing Figure 1 (Snapchat) and Figure 2 (Facebook), I found that both companies scored 0 points for requirement 15. Snapchat also scored a 0 for requirement 4, while Facebook scored 4 out of a possible 24 points. Requirement 15 requests the inclusion of data indicating flags that were traced to bots. Similarly, requirement 4 requests data regarding the number of appeals for moderation decisions made by automated systems. These are the only other two requirements in the *Numbers* principle, outside those included in category d) *automated decision-making tools* that request data related to AI and automatic content moderation tools. As previously proven, Facebook and Snapchat scored the lowest in category d) *automated decision-making tools*. Both companies also scored low in the only other two requirements

regarding automatic moderation processes (requirements 4 and 15), providing further evidence of the companies' continued exclusion of data regarding automatic content moderation practices.

D) Companies are Capable of Improving Scores in Category d) *Automated Decision-Making Tools*

Contrasting Facebook and Snapchat, Pornhub scored the highest in category d) *automated decision-making tools* and the lowest in category b) *state involvement,* where Facebook and Snapchat scored the highest. This is illustrated in Figure 3, which was constructed based on the model of Figures 1 and 2. As explained in my methodology section, I have divided Pornhub's published reports into 10 reports, which all cover six-month intervals since their first published report in 2020. As such, Pornhub can receive a maximum score of 10 for each of the requirements listed along the x-axis.

Category b) *state involvement* (requirements 8 to 13 along the x-axis) asks for data revealing the identity of state actors who have submitted content moderation requests, whether the request was made based on breaking the law or community guidelines, and to what extent the company responded to state-made requests. None of this information is included in any of Pornhub's published reports. The only requirement in category b) *state involvement* for which Pornhub received points was requirement 8, which asks for the number of requests made by state actors. In addition, Pornhub received its highest score for category d) *automated decision-making*, where they received maximum points for requirements 17, 18, and 23 which request data regarding how automatic tools are used, for which types of content they are implemented, and in which cross-platform automatic content moderation programs they partake in. As such, they scored highest in the category where Facebook and Snapchat received the

fewest points, indicating that it is possible to improve scores in category d) *automated decision-making.*

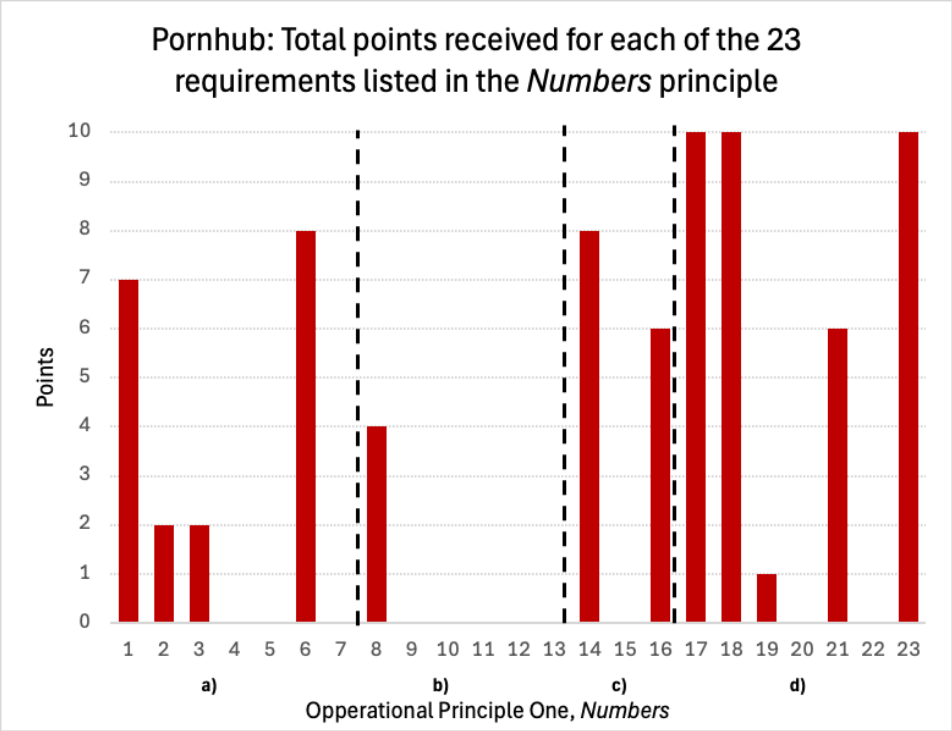

Fig. 3: Comparing all of Pornhub's reports with the 23 requirements in the *Numbers* Principle

E) Increasing the Amount of Information in Transparency Reports Does Not Guarantee Higher Rates of Transparency

As discussed in my methodology section, Facebook releases multiple reports under seven thematic categories. Each category includes subcategories of reports, and all of these are published following different schedules. Some are published monthly, and others are published biannually or annually. Therefore, I constructed my own time parameters for this study that align with Snapchat's biannual publication structure. All information that Facebook has released in reports falling within one of these biannually constructed time parameters is included in the 24 reports I have compiled for this study. Facebook releases a different number of reports each year,

and as such, it releases a different amount of information. Figure 4 compares the number of reports Facebook has released each year and the transparency score it has received in this study.

Facebook released only two reports in 2013, the first year it published a transparency report. The last reports published by Facebook, covering the period from *July to December 2024,* include 26 reports. This is a 1200% increase in the number of reports released. However, in 2013, Facebook scored only 1 out of 24 possible points, and in 2024, it scored 9 out of 24 possible points, representing an 800% increase. Figure 4 shows that the number of reports and the score Facebook received increased consistently until 2024, when the number of reports drastically decreased. This is likely due to not all reports covering the second half of 2024 having been released yet. As such, examining the data point from the first half of 2024, I found a 1,500% increase in the number of reports published since 2013. The score received for this report also showed an 800% rise since 2013, 700% less than the increase in the number of reports published. This shows a disproportionate increase in the amount of information published.

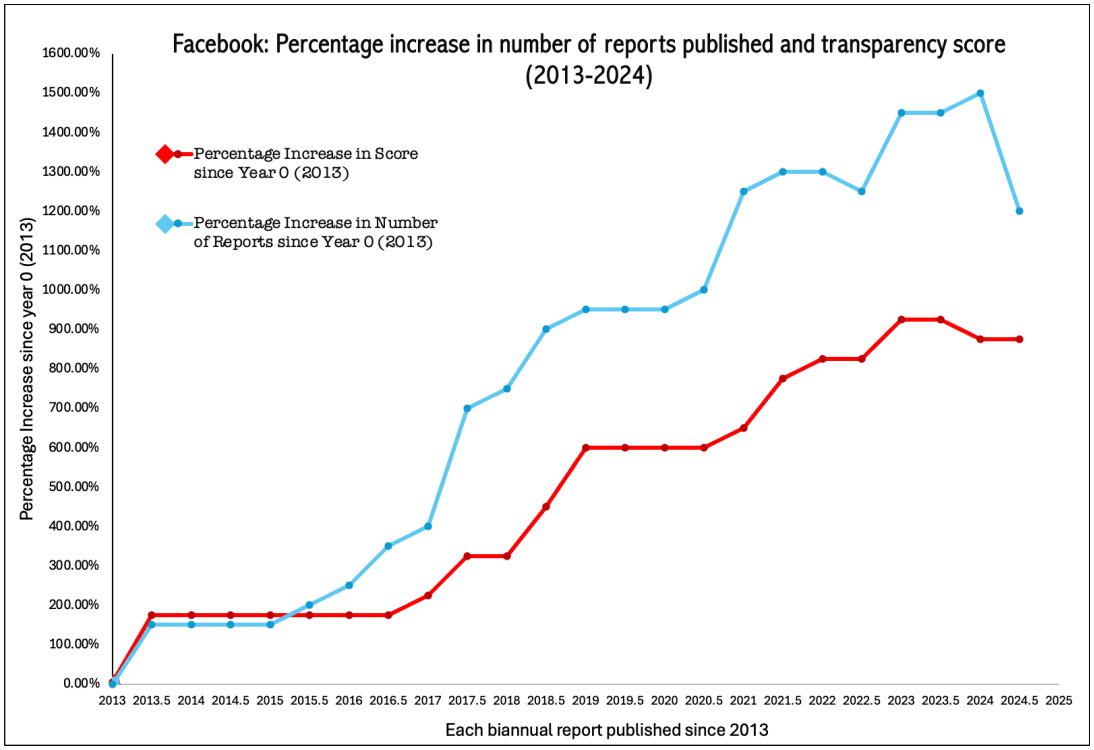

Fig. 4: Comparing the percentage increase in the number of reports published by Facebook and the transparency score of each report

Snapchat has published one report every six months since releasing the first transparency reports in 2014. As such, the number of reports published each year has remained constant. However, the amount of information included in each report has changed. In 2014, the report consisted of only 360 words, but their latest report contained approximately 10,875 words, representing a 2,921% increase. Simultaneously, their first report scored 2.5 points, and their last 12 points, an increase of 380%. Therefore, since the first year they published a report, their transparency score has increased by 2,541% less than the amount of information they have published. It is worth noting that the drastic increase in word count began in 2021, the first year that Snapchat began dividing its reporting into 53 country-specific reports and one global report. As such, the increase is likely not as severe as many words are repeated for each country-specific report. Yet, it is still true that there is a rapidly growing gap between their transparency score and

the amount of information they publish, revealing that more information does not guarantee an equally high score.

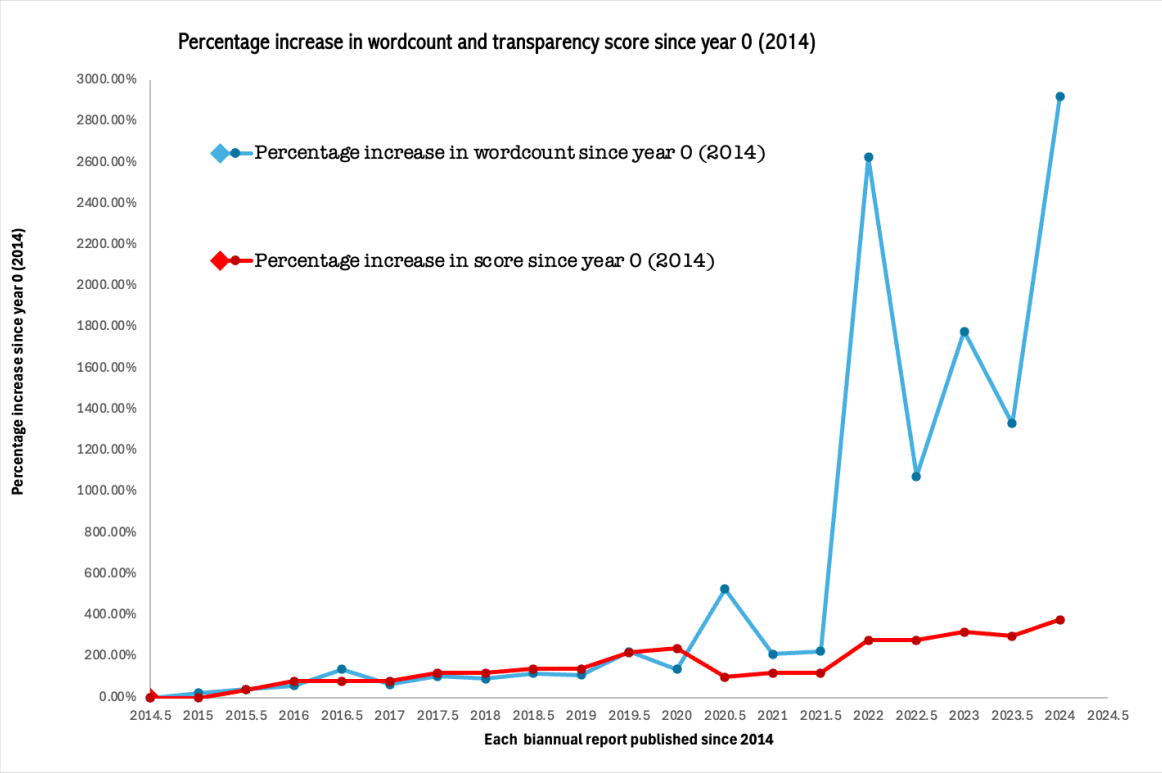

Fig. 5: Comparing Snapchat's percentage increase of each report's word count and transparency score

F) Increasing Information in the Reports Can Increase Transparency

Figure 6, which depicts the same relationship as described in Figure 5 but for Pornhub, shows a different story than Snapchat and Facebook. Here, the red line (transparency score) has increased more than the blue line (word count). For their latest report, published in 2024, their score increased by 186% since 2020, while their word count increased by only 110%, resulting in a 76% higher increase in score compared to the word count. Consequently, as they have published more information, their score has increased at an even greater rate. This demonstrates that it is possible to add information to the reports that targets explicitly areas where compliance with the *Santa Clara Principles 2.0* has been low.

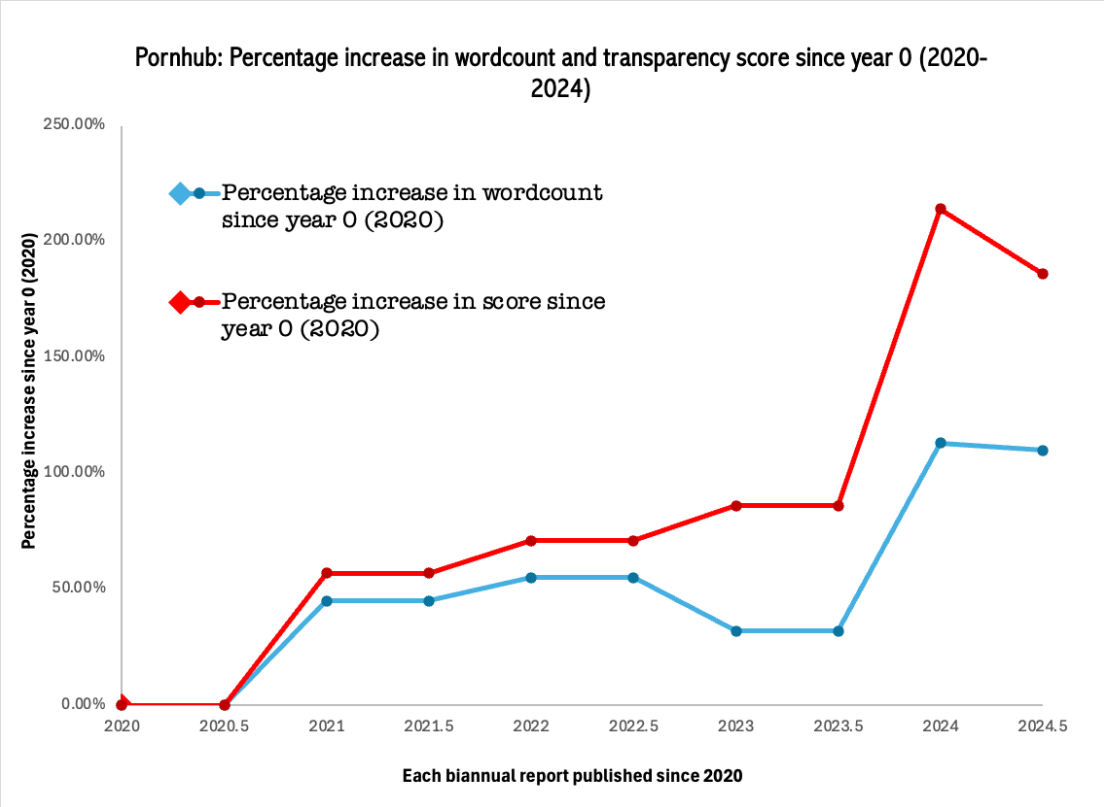

Fig. 6: Comparing Pornhub's percentage increase of each report's word count and transparency score

V. Discussion

The Convention on the Rights of the Child obliges member states to prevent the sexual exploitation of children.[97] Despite this, the National Center for Missing and Exploited Children (NCMEC) has reported an increasing number of cyber tips for child sexual abuse material (CSAM), including child sex trafficking material (CSTM) online.[98] Researchers have argued that a significant factor in the inability to stop the sexual abuse of children online is a lack of content moderation;[99] a process which in the United States has been outsourced to the tech firms who own the platforms used for the distribution of CSTM. However, the same companies increase profits by lowering the amount of content moderation they engage in. Therefore, monitoring is essential to ensure that the firms are not engaging in shirking. In other words, monitoring plays a crucial role in corporate responsibility in preventing the spread of CSTM on online platforms.

To lower the boundaries for monitoring, Facebook, Snapchat, and Pornhub have released transparency reports. However, when comparing the reports to the *Santa Clara Principles 2.0,* it is clear that these reports have significant issues. As can be seen in Figures 1 and 2, Facebook and Snapchat scored continuously low in category d) *automated decision-making tools.* These are the most essential tools in the fight against CSTM as they allow companies to scan numerous posts simultaneously and can identify content that has been flagged as CSTM on other websites.[100] On the other hand, Pornhub scored the highest in the same category. This reveals that there is no lack of technical capability or capacity to present this information. Instead, it is an active choice to exclude data in this area. Similarly, an argument can be made that becuase Snapchat and Facebook scored the highest in category b) *state involvement* (the category where Pornhub preformed the worst), it is possible for Pornhub to imporve its scores in this area too. In

[97] "The Convention on the Rights of the Child," 1-15.
[98] Blegen, 1269.
[99] Veldhuizen-Ochodnicanová, presentation.
[100] Cho and Zhu, 1.

other words, the discrepancy between company performance in each area of transparency reporting displays that there is technological capability and capacity to increase scores in these categories. Yet, the companies must voluntarily choose to do so.

Figures 5 and 6 also show that Facebook and Snapchat have increased the amount of information they publish in their transparency reports by 1200% and 2921%. However, their transparency score has not increased at a similar rate. This suggests that simply increasing the amount of information they publish is insufficient to enhance the quantity of *meaningful* information available. Instead, it further puts weight on the question of "how transparent transparency reports actually are?"[101] and if they are published as proof of the company's transparent behavior or merely to increase the appearance of transparency. Some researchers have found that potential advertisers rely on transparency reports to assess the website's environment for users and its approach to brand safety. Therefore, increasing the amount of information in the reports may be beneficial from a profit-maximizing standpoint for the companies. However, it matters what type of information is included in the reports. As previously discussed, the "Streisand Effect" may deter potential advertisers who study the transparency reports by drawing their attention to the problem of CSTM.[102] Consequently, Facebook and Snapchat may refrain from including data that would highlight the presence of CSTM on their platforms. As automatic content moderation systems are widely used in the filtering of CSTM, it may, as such, be beneficial to exclude data about these.

Simultaneously, as discussed earlier, controversial content that does not align with the viewer's moral ideology may keep users more engaged on the platform. As such, they will keep scrolling and be exposed to more advertisements.[103] Therefore, it is also in the company's interest

[101] Makhortykh and Urman, 1.
[102] James Grimmelman. "The Virtues of Moderation," *Yale Journal of Law & Technology* 42 (2015): 66.
[103] Mochon and Schwartz, 1.

to minimize content moderation while enhancing the perception of transparency through its transparency reports. Figures 5 (Snapchat) and 6 (Facebook) show that both companies have rapidly increased the amount of information in their transparency reports, without their transparency score increasing at a similar rate. In other words, they are expanding the perception of transparency of content moderation, which attracts advertisers to their platforms, while continuing to minimize external monitoring mechanisms of their content moderation systems.

Figure 6 (Pornhub) does not mirror this discrepancy in changes in scores and information released over time. Instead, as Pornhub has published more information, its transparency score has increased at an even greater rate. As such, it is clear that the additional information is *meaningful* and has been added to increase transparency regarding their content moderation practices, rather than to enhance the appearance of corporate transparency. Again, this demonstrates that it is possible to modify the reports to enhance the actual transparency of the content moderation process, thereby improving monitoring capabilities. This would strengthen accountability mechanisms for content moderation, helping to prevent shirking and the spread of CSTM online, and thereby increase the protection of children's rights.

VI. Conclusion

As I write this thesis in the Spring of 2025, current events are essential to understanding the significance of the project. At the beginning of the year, Meta (Facebook) announced that they would be cutting back on automatic content moderation and laying off thousands of human moderators in favor of a new content moderation program called "Community Notes" that has already been implemented by YouTube and X. This moderation system would minimize internal moderation and rely on users reporting and editing content on the company's platforms. The "Community Note" system has been highly criticized, and Mark Zuckerberg has said that it would increase the amount of disinformation and harmful content on the platform, but reduce censorship and improve freedom of speech.[104] However, the company intends to keep using automatic content moderation to scan for illegal posts such as CSAM and CSTM.[105] As human moderators are being dismissed and Facebook is showing clear signs of prioritizing free speech over content moderation, these automatic content moderation tools are becoming increasingly critical in the fight against CSTM on their platform. In this study, I have identified a concerning trend in which Facebook and Snapchat consistently refrain from disclosing information about their internal automatic content moderation practices. As these tools now become the only remaining company-implemented content moderators, publishing meaningful information about them becomes increasingly significant. Facebook has effectively removed its liability for moderation as independent users will conduct it. As such, all corporate liability for moderation will ultimately depend on the automatic content moderation systems. Therefore, to hold

[104] Mike Isaac and Theodore Schleifer. "Meta Says It Will End Its Fact-Checking Program on Social Media Posts." *New York Times,* January 7, 2025.
[105] Clare Duffy. "Meta Gets Rid of Fact Checkers. Zuckerberg Acknowledged More Harmful Content Will Appear on the Platforms Now." *CNN, January 7, 2025.*

Facebook accountable for its efforts to prevent CSTM on its platform, transparency regarding the implementation of AI and automatic content moderation systems must increase.

Bibliography

Ardigó, Iñaki Albisu. “Corporate transparency: Overview of issues and literature review,” *Transparency International,* April 16, 2019. https://www.jstor.org/stable/resrep20480.

Blegen, Elise N. “Sex, Drugs, and Rock & Roll: Effectively and Equitably Moderating Vice and Illegal Content Online." *Vanderbilt Law Review* 76 no. 4 (2023): 1265-1298.

Bouché, Vanessa. *A Report on the Use of Technology to Recruit, Groom, and Sell Domestic Minor Sex Trafficking Victims:* Thorn (2015). https://www.thorn.org/wp-content/uploads/2015/02/Survivor_Survey_r5.pdf.

Bouché, Vanessa. *Survivor Insights: the Role of Technology in Domestic Minor Sex Trafficking:* Thorn (2018). https://www.thorn.org/wp-content/uploads/2018/06/Thorn_Survivor_Insights_061118.pdf

Campagna, Daniel S and Proffenberger, Donald D. *The sexual trafficking in children: an investigation of the child sex trade,* Dover, MA, Auburn House, 1988.

Cho, Clare and Ling Zhu. “Social Media: Content Dissemination and Moderation Practices.” *Congress.gov* (2025): 1. https://crsreports.congress.gov/product/pdf/R/R46662.

Cobbe, Jennifer and Rozgonyi, Krisztina and Sekwenz, Marie-Therese and Wagner, Ben. “Regulating Transparency?: Facebook, Twitter and the German Network Enforcement Act,” *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, New York: Association for Computing Machinery, (2020): 261–271. https://doi.org/10.1145/3351095.3372856.

“CyberTipline 2023 Report,” National Center for Missing and Exploited Children. Accessed April 17, 2025, https://www.missingkids.org/cybertiplinedata.

Duffy, Clare. “Meta Gets Rid of Fact Checkers. Zuckerberg Acknowledged More Harmful Content Will Appear on the Platforms Now.” *CNN, January 7, 2025.* https://www.cnn.com/2025/01/07/tech/meta-censorship-moderation/index.html.

Farvaque, Etienne and Refait-Alexandre, Catherine and Saïdane, Dhafer. “Corporate Disclosure: A Review of its (Direct and Indirect) Benefits and Costs,” *International Economics* 128, no. 4 (2011): 5-32. https://doi.org/10.3917/ecoi.128.0005.

Franks, Mary A, "Justice Beyond Dispute," *Harvard Law Review* 131 (2018): 1374-1397. https://www.jstor.org/stable/i40206550

Grimmelman, James. "The Virtues of Moderation," *Yale Journal of Law & Technology* 42 (2015): 43-108. https://scholarship.law.cornell.edu/facpub/1486/.

Hackler, Jacqueline. "Inconsistencies in Combatting the Sex Trafficking of Minors: Backpage's Deceptive Business Practices Should Not Be Immune from State Law Claims." *Seattle University Law Review* 40, (2017): 1109. https://digitalcommons.law.seattleu.edu/sulr/vol40/iss3/7/.

Halbac, Claudia. Lecture given in American Economy Class at Columbia University, New York, NY, October 15, 2024.

Haney, Rachel. "When 'Safe at home' is Not Safe," *Scitech Lawyer* 17, no. 4. (2021): 22-27. https://www.proquest.com/docview/2585480688?pq-origsite=gscholar&fromopenview=true&sourcetype=Scholarly%20Journals.

"How Facebook and Instagram Became Market Places for Child Sex Trafficking," *Guardian,* Apr. 27, 2023. https://www.theguardian.com/news/2023/apr/27/how-facebook-and-instagram-became-marketplaces-for-child-sex-trafficking.

Isaac, Mike and Theodore Schleifer. "Meta Says It Will End Its Fact-Checking Program on Social Media Posts." *New York Times,* January 7, 2025. https://www.nytimes.com/live/2025/01/07/business/meta-fact-checking.

Kowalski, Sandra E. "Holding Internet Advertising Providers Accountable For Sex Trafficking: Implements to Criminal Prosecution and a Proposed Response" *Public Interest Law Journal* 27 (2018): 99-132.

Kristof, Nicholas. "The Children of Pornhub." *New York Times*, Dec. 4, 2020. https://www.nytimes.com/2020/12/04/opinion/sunday/pornhub-rape-trafficking.html.

Levine, Brian N. "Increasing the Efficiency of Investigations of Online Child Sexual Exploitation." University of Massachusetts Amherst, 2022.

Lorenzo-Dus, Nuria. *Digital Grooming: Discourses of Manipulation and Cyber-Crime,* Oxford: Oxford University Press, 2022.

Makhortykh, Mykola and Urman, Aleksandra. "How Transparent are Transparency Reports? Comparative Analysis of Transparency Reporting Across Online Platforms,"

Telecommunications Policy 47. no. 3. (2023): 1-15. https://doi.org/10.1016/j.telpol.2022.102477.

Mochon, Daniel and Janet Schwartz. "The Confrontation Effect: When Users Engage More with Ideology-Inconsistent Content Online." *Organizational Behavior and Human Decision Processes* 185 (2024): 1, https://doi.org/10.1016/j.obhdp.2024.104366.

Mungiu-Pippidi, Alina. "Measuring real (de facto) transparency by a new index" *European Research Centre for Anti-Corruption and State-Building,* (2021): 1-18.

Rosenzweig, Paul and Katie Stoughton. "Toward Greater Content Moderation Transparency Reporting." *Lawfare (*2022). https://www.lawfaremedia.org/article/toward-greater-content-moderation-transparency-reporting.

"The Santa Clara Principles On Transparency and Accountability in Content Moderation," Santa Clara Principles. Accessed November 22, 2024, https://santaclaraprinciples.org/.

"Transparency Reports," Meta. Accessed November 22, 2023, https://transparency.meta.com/reports/.

Uittis, Beatriz S. *Sex Trafficking of Children Online: Modern Slavery in Cyberspace.* Lanham: Rowman & Littlefield, 2022.

United Nations Human Rights Committee (HRC). "CCPR General Comment No. 17: Article 24 (Rights of the Child)" (1989): 1-3.

United Nations Human Rights Office of the High Commissioner. "Human Rights and Human Trafficking" *Fact Sheet No. 36* (2014): 1-59.

United Nations, "Convention on the Elimination of All Forms of Discrimination Against Women," *Treaty Series 1249* (1988).

United Nations, "The Convention on the Rights of the Child." *Treaty Series 1577* (1989): 1-15.

United Nations International Children's Emergency Fund, "Becoming the leading voice: 2006–2020." Accessed April 17, 2025. https://www.unicef.org/stories/learning-experience-20062015.

United Nations International Children's Emergency Fund, "Bringing the world together: 1989–2005." Accessed April 17, 2025. https://www.unicef.org/stories/learning-experience-19892005.

United Nations International Children's Emergency Fund, "Children's Rights and Business Principles," (2012): 1-21.

United Nations International Children's Emergency Fund, "Child Rights and Why They Matter: Every Right, For Every Child." Accessed April 17, 2025. https://www.unicef.org/child-rights-convention/child-rights-why-they-matter.

United Nations International Children's Emergency Fund, "Emerging from the ashes of war: 1946–1979." Accessed April 17, 2025. https://www.unicef.org/stories/learning-experience-19461979.

United Nations International Children's Emergency Fund, "Moving with the times: 1980–1988." Accessed April 17, 2025. https://www.unicef.org/stories/learning-experience-19801988.

United Nations (General Assembly). "International Covenant on Civil and Political Rights." *Treaty Series* 999 (1966): 171.

United Nations, *Protocol to Prevent, Suppress and Punish Trafficking in Persons, Especially Women and Children, supplementing the United Nations Convention against Transnational Organized Crime*, 2237 U.N.T.S. 319 (2000).

Veldhuizen-Ochodnicanová, Eva. "Technology-facilitated Child Abuse, Exploitation and Trafficking" April 8, 2024, Johns Hopkins School of Advanced International Studies, Baltimore, Maryland, United States. https://bipr.jhu.edu/events/4793-The-Human-Trafficking-Initiative-Series--Technology-facilitated-Child-Abuse-Expl.cfm.

"What is A Transparency Report," Trust and Safety Professional Association. Accessed November 22, 2024. https://www.tspa.org/curriculum/ts-fundamentals/transparency-report/what-is-a-transparency-report/#:~:text=Transparency%20reporting%20can%20be%20important,for%20openness%20with%20the%20public.

Appendix A: Table Displaying Snapchat's Transparency Score for Each Published Report

	1.	**2.**	**3.**	**4.**	**5.**	**6.**	**7.**	**8.**	**9.**	**10.**	**11.**	**12.**	**13.**	**14.**	**15.**	**16.**	**17.**	**18.**	**19.**	**20.**
1**	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
2	0.5	0	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0	0	0	0.5	0.5	1	1	1
3	0.5	0	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0	0	0	0.5	0.5	1	1	1
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
7	0	0	0	0	0	0	0	0	0	0	0	0	0.5	1	1	1	1	1	1	1
8	0.5	1	1	1	1	1	1	1	1	1	1	1	0	0	0	1	1	1	1	1
9	0	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0	0	0	0.5	0.5	0.5	0.5	0.5
10	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0	0	0	0.5	0.5	0.5	0.5	0.5
11	0	0	0	1	1	1	1	1	1	1	1	1	0	0	0	1	1	1	1	1
12	0	0	0	0	0	0	0.5	0.5	0.5	0.5	0.5	0	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0.5	0.5	1	1	1	0	0	0	0	0	0	0	0	0
14	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	1	1	1	1	1	1	1	1	1	1
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	0	0	0	0	0	0	0	0	0	0	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.5
18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1

Fig. 7 (Snapchat's Transparency reports compared to Operational Principle One, *Numbers* (SCP 2.0))

** Operational Principle One, *Numbers*, data point a) to w) as explained in Appendix H (valid for all rows in Fig. 7)

* Snapchat's transparency reports as listed in Appendix F (valid for all columns in Fig. 7)

Appendix B: Table Displaying Pornhub's Transparency Score for Each Published Report

	1.*	**2.**	**3.**	**4.**	**5.**	**6.**	**7.**	**8.**	**9.**	**10.**
1**	0.5	0.5	0.5	0.5	0.5	0.5	1	1	1	1
2	0	0	0	0	0	0	0	0	1	1
3	0	0	0	0	0	0	0	0	1	1
4	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0
6	0	0	1	1	1	1	1	1	1	1
7	0	0	0	0	0	0	0	0	0	0
8	1	1	0	0	0	0	0	0	1	1
9	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0
14	0	0	1	1	1	1	1	1	1	1
15	0	0	0	0	0	0	0	0	0	0
16	0	0	1	1	0.5	0.5	0.5	0.5	1	1
17	1	1	1	1	1	1	1	1	1	1
18	1	1	1	1	1	1	1	1	1	1
19	0	0	0	0	0	0	0	0	1	0
20	0	0	0	0	0	0	0	0	0	0
21	0	0	0	0	1	1	1	1	1	1
22	0	0	0	0	0	0	0	0	0	0
23	1	1	1	1	1	1	1	1	1	1

Fig. 8 (Pornhub's Transparency reports compared to Operational Principle One, *Numbers* (SCP 2.0))

** Operational Principle One, *Numbers*, data point a) to w) as explained in Appendix H (valid for all rows in Fig. 8)

* Pornhub's transparency reports as listed in Appendix F (valid for all columns in Fig. 8)

Appendix C: Table Displaying Facebook's Transparency Score for Each Published Report

	1.*	**2.**	**3.**	**4.**	**5.**	**6.**	**7.**	**8.**	**9.**	**10.**	**11.**	**12.**	**13.**	**14.**	**15.**	**16.**	**17.**	**18.**	**19.**	**20.**	**21.**	**22.**	**23.**	**24.**
1**	0	0	0	0	0	0	0	0	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
2	0	0	0	0	0	0	0	0	0	0	0	0	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
3	0	0	0	0	0	0	0	0	0	0	0	0	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
5	0	0	0	0	0	0	0	0	0	0	0	0	0.5	0.5	0.5	0.5	0.5	0.5	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1
7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
9	0	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
10	0	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	0	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	1	1	1	1	1	1
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	0	0	0	0	0	0	0	0	0	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.5	0.5	0.5	0.5	0.5	0.5
18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.5	0.5	0.5	0.5
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.5	0.5	0.5	0.5
20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.5	0.5	0.5	0.5	0.5	0.5	0.5
21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.5
22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.5	0.5	0.5	0.5	0.5	0	0

Fig. 9 (Facebook's Transparency reports compared to Operational Principle One, *Numbers* (SCP 2.0))

** Operational Principle One, *Numbers*, data point a) to w) as explained in Appendix H (valid for all rows in Fig. 9)

* Facebook's transparency reports as listed in Appendix F (valid for all columns in Fig. 9)

Appendix D: Explanations of All Half-Points Recived By Each Company

Row	Column	Reason for points
2 & 3	1, 3-12, 16-17	Data is only provided in the case of The *Digital Millennium Copyright Act* (DMCA)
7	13	Data is only provided regarding spreading false information about the COVID-19 pandemic.
8 - 10	1 - 12, 16 - 20	Data is only provided for the United States of America.
12	7-11	It does not specify which specific laws/rules the account or content violated.
13	7-8	Data is only reported when the content or account violates local laws.
14	1-10	Data is only reported when a state actor makes the request.
16	11-20	It does not reveal the source of the flag submission.
17	20	It does not explain how they are used in relation to human oversight.

Fig. 10 (Explanation for all half points awarded to Snapchat in Fig. 7)

Row	Column	Reason for points
1	1-3	It does not include data showing the number of accounts actions were taken against.
1	6	It does not report the total number of pieces of content removed.
1	8	It does not include actions taken for reports made by state governments
16	3-5, 7	The report does not identify the source of the submitted flag.

Fig. 11 (Explanation for all half points awarded to Pornhub in Fig. 8)

Row	Column	Reason for points
1	9-11	It does not report data on content removed under the following categories: Violence and Incitement, Suicide and Self-Injury, Restricted Goods and Services, Bullying and Harassment, and Child Endangerment. It does not distinguish between removed content and suspended accounts.
1	12	It does not report data on content removed under the following categories: Violence and Incitement and Suicide and Self-Injury. It does not distinguish between removed content and suspended accounts.
1	13-17	It does not report data on content removed under the following categories: Violence and Incitement. It does not distinguish between removed content and suspended accounts.
1	18-21	It does not differentiate between removed content and suspended accounts.

1	22-24	Only in the case of European Economic Area countries is the data separated between removed content and suspended accounts.
2	15-17	It does not provide data for the following categories: Fake Accounts & Violence and Incitement.
2	18-24	The report does not provide data for the following categories: Fake Accounts.
3	13-15	The report does not provide data for the following categories: Fake Accounts, Child Endangerment, and Violence and Incitement.
3	16-17	It does not provide data for the following categories: Fake Accounts and Violence and Incitement.
3	18-19	The report does not provide data for the following categories: Fake Accounts.
3	20	This was only reported in the case of EU Terrorist Content.
3	21-23	This was only reported in the case of European Economic Area countries.
4	17-24	This was only reported when automated detection tools were used to prevent the spread of CSAM and at least one EU user was involved.
5	13-16	The report does not provide data for the following categories: Fake Accounts, Child Endangerment, and Violence and Incitement.
5	17	It does not provide data for the following categories: Fake Accounts and Violence and Incitement.
5	18	It does not provide data for the following categories: Fake Accounts.
9 - 10	2-21	Data is only reported in the case of the United States of America.
9 - 10	22-24	Data is only reported in the case of the United States of America and Turkey.
12	2-24	The report does not specify which local laws the content violated.
14	10-11	It does not report data reflecting the number of flags received for the following categories: Bullying and Harassment, Child Endangerment, Restricted Goods and Services, Suicide and Self-Injury, and Violence and Incitement.
14	12	It does not report data reflecting the number of flags received for the following categories: Child Endangerment, Suicide and Self-Injury, and Violence and Incitement.
14	13-16	It does not report data reflecting the number of flags received for the following categories: Child Endangerment & Violence and Incitement.
14	17-18	Data reflecting the number of flags received for all categories is only reported for India.
16	10-24	The report does not reveal the source of the submitted flag.
17	19	Data is reported only in the case of NetzDG (Germany) and the EU Terrorist Content Online Transparency Report.
17	20	Data is reported only in the case of NetzDG (Germany).
17 - 19	21-24	Data only reported in the case of European Economic Area countries.

20	18-22	Data is only provided for automatic detection tools used to combat CSAM.
20	23-24	Data is only reported in the case of European Economic Area countries and CSAM.
21	24	The extent to which there is human oversight is described in very vague terms, but only in the case of automatic moderation tools implemented in Turkey.
23	18-19	It is only mentioned in the case of CSAM.
23	20	It is only mentioned in the case of CSAM and the EU Terrorist Content Online Transparency Report.
23	21-22	It is only mentioned in the case of Korea and the EU Terrorist Content Online Transparency Report.

Fig. 12 (Explanation for all half points awarded to Facebook in Fig. 9)

Appendix E: Summary of Scores Recived for Each Report Published By the Three Companies

Report	Total Score
Nov. 1, 2014 - Feb. 28, 2015	2.5/23 ≈ 11%
Jan. 1, 2015 - June 30, 2015	2.5/23 ≈ 11%
July 1, 2015 - Dec. 31, 2015	4/23 ≈ 17%
Jan. 1, 2016 - June 30, 2016	4.5/23 ≈ 20%
July 1, 2016 - Dec. 31, 2016	4.5/23 ≈ 20%
Jan. 1, 2017 - June 31, 2017	4.5/23 ≈ 20%
July 1, 2017 - Dec. 31, 2017	5.5/23 ≈ 24%
Jan. 1, 2018 - June 30, 2018	5.5/23 ≈ 24%
July 1, 2018 - Dec. 31, 2018	6/23 ≈ 26%
Jan. 1, 2019 - June 30, 2019	6/23 ≈ 26%
July 1, 2019 - Dec. 31, 2019	9/23 ≈ 39%
Jan. 1, 2020 - June 30, 2020	8.5/23 ≈ 37%
July 1, 2020 - Dec. 31, 2020	5/23 ≈ 22%
Jan. 1, 2021 - June 30, 2021	5.5/23 ≈ 24%
July 1, 2021 - Dec. 31, 2021	5.5/23 ≈ 24%
Jan. 1, 2022 - June 30, 2022	9.5/23 ≈ 41%
July 1, 2022 - Dec. 31, 2022	9.5/23 ≈ 41%
Jan. 1, 2023 - June 30, 2023	10.5/23 ≈ 46%
July 1, 2023 - Dec. 31, 2023	10.5/23 ≈ 46%
Jan. 1, 2024 - June 30, 2024	12/23 ≈ 52%

Fig. 13 (Total points received by each published Snapchat transparency report)

Report	Total Score
2020 Transparency Report	4.5/23 ≈ 20%
2021 Transparency Report	6.5/23 ≈ 28%
2022 Transparency Report	7/23 ≈ 30%
2023 Transparency Report (First Half)	7.5/23 ≈ 33%
2023 Transparency Report (Second Half)	7.5/23 ≈ 33%
EU DSA Transparency Report - August 2024	10.5/23 ≈ 46%
2024 Transparency Report (First Half)	7.5/23 ≈ 33%

Fig. 14 (Total points received by each published Pornhub transparency report)

Report	Total Score
Jan. - June, 2013	1/23 ≈ 4%
July - Dec. 2013	2.5/23 ≈ 11%
Jan. - June, 2014	2.5/23 ≈ 11%
July - Dec. 2014	2.5/23 ≈ 11%
Jan. - June, 2015	2.5/23 ≈ 11%
July - Dec. 2015	2.5/23 ≈ 11%
Jan. - June, 2016	2.5/23 ≈ 11%
July - Dec. 2016	2.5/23 ≈ 11%
Jan. - June, 2017	3/23 ≈ 13%
July - Dec. 2017	4/23 ≈ 17%
Jan. - June, 2018	4/23 ≈ 17%
July - Dec. 2018	5/23 ≈ 22%
Jan. - June, 2019	6.5/23 ≈ 28%
July - Dec. 2019	6.5/23 ≈ 28%
Jan. - June, 2020	6.5/23 ≈ 28%
July - Dec. 2020	6.5/23 ≈ 28%
Jan. - June, 2021	7/23 ≈ 30%
July - Dec. 2021	8/23 ≈ 35%
Jan. - June, 2022	8.5/23 ≈ 37%
July - Dec. 2022	8.5/23 ≈ 37%
Jan. - June, 2023	9.5/23 ≈ 41%
July - Dec. 2023	9.5/23 ≈ 41%
Jan. - June, 2024	9/23 ≈ 39%
July - Dec. 2024	9/23 ≈ 39%

Fig. 15 (Total points received by each published Facebook transparency report)

Appendix F: List of All Reports Included in This Study

1. Nov. 1, 2014 - Feb. 28, 2014
2. Jan. 1, 2015 - June 30, 2015
3. July 1, 2015 - Dec. 31, 2015
4. Jan. 1, 2016 - June 30, 2016
5. July 1, 2016 - Dec. 31, 2016
6. Jan. 1, 2017 - June 31, 2017
7. July 1, 2017 - Dec. 31, 2017
8. Jan. 1, 2018 - June 30, 2018
9. July 1, 2018 - Dec. 31, 2018
10. Jan. 1, 2019 - June 30, 2019
11. July 1, 2019 - Dec. 31, 2019
12. Jan. 1, 2020 - June 30, 2020
13. July 1, 2020 - Dec. 31, 2020
14. Jan. 1, 2021 - June 30, 2021
15. July 1, 2021 - Dec. 31, 2021
16. Jan. 1, 2022 - June 30, 2022
17. July 1, 2022 - Dec. 31, 2022
18. Jan. 1, 2023 - June 30, 2023
19. July 1, 2023 - Dec. 31, 2023
20. Jan. 1, 2024 - June 21, 2024

Fig. 16 (all published Snapchat transparency reports)

1. 2020 Transparency Report (First Half)
2. 2020 Transparency Report (Second Half)
3. 2021 Transparency Report (First Half)
4. 2021 Transparency Report (Second Half)
5. 2022 Transparency Report (First Half)
6. 2022 Transparency Report (Second Half)
7. 2023 Transparency Report (First Half)
8. 2023 Transparency Report (Second Half)
9. 2024 Transparency Report (First Half) & EU DSA Transparency Report - August 2024
10. 2024 Transparency Report (Second Half) and EU DSA Transparency Report - February 2025

Fig. 17 (all published Pornhub transparency reports)

1. Jan. - June, 2013
2. July - Dec. 2013
3. Jan. - June, 2014
4. July - Dec. 2014
5. Jan. - June, 2015
6. July - Dec. 2015
7. Jan. - June, 2016
8. July - Dec. 2016
9. Jan. - June, 2017
10. July - Dec. 2017
11. Jan. - June, 2018
12. July - Dec. 2018
13. Jan. - June, 2019
14. July - Dec. 2019
15. Jan. - June, 2020
16. July - Dec. 2020
17. Jan. - June, 2021
18. July - Dec. 2021
19. Jan. - June, 2022
20. July - Dec. 2022
21. Jan. - June, 2023
22. July - Dec. 2023
23. Jan. - June, 2024
24. July - Dec. 2024

Fig. 18 (all constructed Meta transparency reports)

Appendix G: List Explaining Meta's Separation of Thematic Reports

Meta's Seven Areas of Transparency Reporting:

1. *Community Standards Enforcement Report*
2. *Intellectual Property*
3. *Government Requests for User Data*
4. *Content Restrictions Based on Local Laws*
5. *Internet Disruptions*
6. *Widely Viewed Content Report*
7. *Regulatory and Other Transparency Reports*

Each Report I have studied that is Published Under each of the Seven Areas Listed Above:

1. *Community Standards Enforcement Report* (published quarterly starting October 2017)
    a. Adult Nudity and Sexual Activity
    b. Bullying and Harassment
    c. Child Endangerment: Nudity and Physical Abuse and Sexual Exploitation
    d. Dangerous Organizations: Terrorism and Organized Hate
    e. Fake Accounts
    f. Hate Speech
    g. Restricted Goods and Services
    h. Spam
    i. Suicide and Self-Injury
    j. Violence and Incitement
    k. Violent and Graphic Content
2. *Intellectual Property*
    a. Notice and takedown (Data provided from 2017-2024)
    b. Proactive enforcement (Data provided from 2020-2024)
    c. How we protect intellectual property (IP) rights
3. *Government Requests for User Data*
    a. Global overview (Jan 13 - June 2024)
    b. Data types
    c. Preservation requests (Jan 2016 - June 2024)
    d. Requests by country (Jan 13 - June 2024)
    e. Further asked questions
4. *Content Restrictions Based on Local Laws*
    a. Overview (July 2013 - June 2024)
    b. Restrictions by country (July 2013 - June 2024)
    c. Global restrictions (July 2013 - June 2024)
    d. Case studies (July 2015 - June 2024)
    e. How we assess reports of content violating local law
5. *Internet Disruptions*

a. Global overview (July 2016 - June 2024)
b. Disruptions by country (July 2016 - June 2024)

6. *Widely Viewed Content Report: What People See on Facebook*
   a. 2021-q2-wvcr
   b. 2021-q3-wvcr
   c. 2021-q4-wvcr
   d. 2022-q1-wvcr
   e. 2022-q2-wvcr
   f. 2022-q3-wvcr
   g. 2022-q4-wvcr
   h. 2023-q1-wvcr
   i. 2023-q2-wvcr
   j. 2023-q3-wvcr
   k. 2023-q4-wvcr
   l. 2024-q1-wvcr
   m. 2024-q2-wvcr
   n. 2024-q3-wvcr
7. *Regulatory and Other Transparency Reports*
   a. EU Digital Markets Act: Meta's Profiling Techniques Description (2024)
   b. EU Digital Markets Act: Meta's Compliance Report (2024)
   c. EU Digital Services Act: Transparency Reports for Very Large Online Platforms (April - Sep. 2024, Oct. 2023 - April 2024, April - Sep. 2023) (Instagram & Facebook)
   d. EU Digital Services Act: Systemic Risk Assessment Results Reports (Sep. 2023 - Aug. 2024) (Instagram & Facebook)
   e. EQ Digital Services Act: Independent Audit Reports (Aug. 2023 - June 2024) (Instagram & Facebook)
   f. EU Digital Services Act: Audit Implementation Reports (Sep. 2024) (Instagram & Facebook)
   g. Digital Services Act - Information on Average Monthly Active Recipients in the European Union (Feb. 2023, Aug. 2023, Feb. 2024, Aug. 2024)
   h. EU Digital Services Act: Post-Election Reports (June 2024)
   i. KoPI-G transparency report (Jan-Jun 2023, Jul-Dec 2022, Jan-Jun 2022, Oct-Dec 2021, Apr-Sep 2021) (Instagram & Facebook)
   j. Korea Transparency Report (2023)
   k. EU CSAM derogation report (2021, 2022, 2023, 2024)
   l. India Monthly Report under the Information Technology (Intermediary Guidelines and Digital Ethics Code) Rules, 2021 (2021, 2022, 2023, 2024, 2025)
   m. NetzDG transparency report (Jan-Jun 2018, Jul-Dec 2018, Jan-Jun 2019, Jul-Dec 2019, Jan-Jun 2020, Jul-Dec 2020, Jan-Jun 2021, Jul-Dec 2021, Jan-Jun 2022, Jul-Dec 2022, Jan-Jun 2023) (Instagram & Facebook)
   n. Türkiye Transparency Report for Content Removal Requests under Law no. 5651 (Jul-Dec 2024, Jan-Jun 2024, Jul-Dec 2023, Jan-Jun 2023, Jul-Dec 2022, Jan-Jun 2022, Mar-Jun 2021, Jul-Dec 2021) (Instagram & Facebook)

o. EU Terrorist Content Online Transparency Report (Jan-Dec 2023, Jun-Dec 2022) (Instagram & Facebook)
p. Norwegian Transparency Act Statement (Jan-Dec 2023)
q. Brazil TSE Election Rules Transparency Reports (Aug. 2024, Dec. 2024)

Appendix H: Operational Principle One, *Numbers*: *The Santa Clara Principles 2.0*

All 23 points that the SCP 2.0 requires companies to include in their transparency reporting divided into the four seperate categories.

1) Total number of pieces of content actioned and accounts suspended.
2) Number of appeals of decisions to action content or suspend accounts.
3) Number (or percentage) of successful appeals that resulted in pieces of content or accounts being reinstated, and the number (or percentage) of unsuccessful appeals and;
4) Number (or percentage) of successful or unsuccessful appeals of content initially flagged by automated detection.
5) Number of posts or accounts reinstated by the company proactively, without any appeal, after recognition that they had been erroneously actioned or suspended.
6) Numbers reflecting enforcement of hate speech policies, by targeted group or characteristic, where apparent, though companies should not collect data on targeted groups for this purpose
7) Numbers related to content removals and restrictions made during crisis periods, such as during the COVID-19 pandemic and periods of violent conflict.

Special reporting requirements apply to decisions made with the involvement of state actors, which should be broken down by country:

8) The number of demands or requests made by state actors for content or accounts to be actioned
9) The identity of the state actor for each request
10) Whether the content was flagged by a court order/judge or other type of state actor
11) The number of demands or requests made by state actors that were actioned and the number of demands or requests that did not result in actioning.
12) Whether the basis of each flag was an alleged breach of the company's rules and policies (and, if so, which rules or policies) or of local law (and, if so, which provisions of local law), or both.
13) Whether the actions taken against content were on the basis of a violation of the company's rules and policies or a violation of local law.

Because there are special concerns that flagging processes will be abused, companies should consider reporting data that will allow users and researchers to assess the frequency of such abuse and the measures a company takes to prevent it. Specific metrics and/or qualitative reporting could be devised to help identify abuse-related trends in particular regional contexts. Companies should consider collecting and reporting the following, broken down by country or region if available:

14) The total number of flags received over a given period of time.
15) The total number of flags traced to bots.
16) The number of posts and accounts flagged, in total, and broken down by
    a) Alleged violation of rules and policies
    b) Source of the flag (state actors, trusted flaggers, users, automation, etc.

Due to the increasing role that automated processes play in content moderation, a comprehensive understanding of companies' processes and systems requires transparency around the use of automated decision-making tools. In addition to the numbers about the use of automation called for above, companies should publish information relating to:

17) When and how automated processes are used (whether alone or with human oversight) when actioning content.
18) The categories and types of content where automated processes are used;
19) The key criteria used by automated processes for making decisions;
20) The confidence/accuracy/success rates of automated processes, including changes over time and differences between languages and content categories;
21) The extent to which there is human oversight over any automated processes, including the ability of users to seek human review of any automated content moderation decisions;
22) The number (or percentage) of successful and unsuccessful appeals when the content or account was first flagged by automated detection, broken down by content format and category of violation;
23) Participation in cross-industry hash-sharing databases or other initiatives and how the company responds to content flagged through such initiatives.[106]

[106] Santa Clara Principles 2.0 Operational Principle One, *Numbers.*