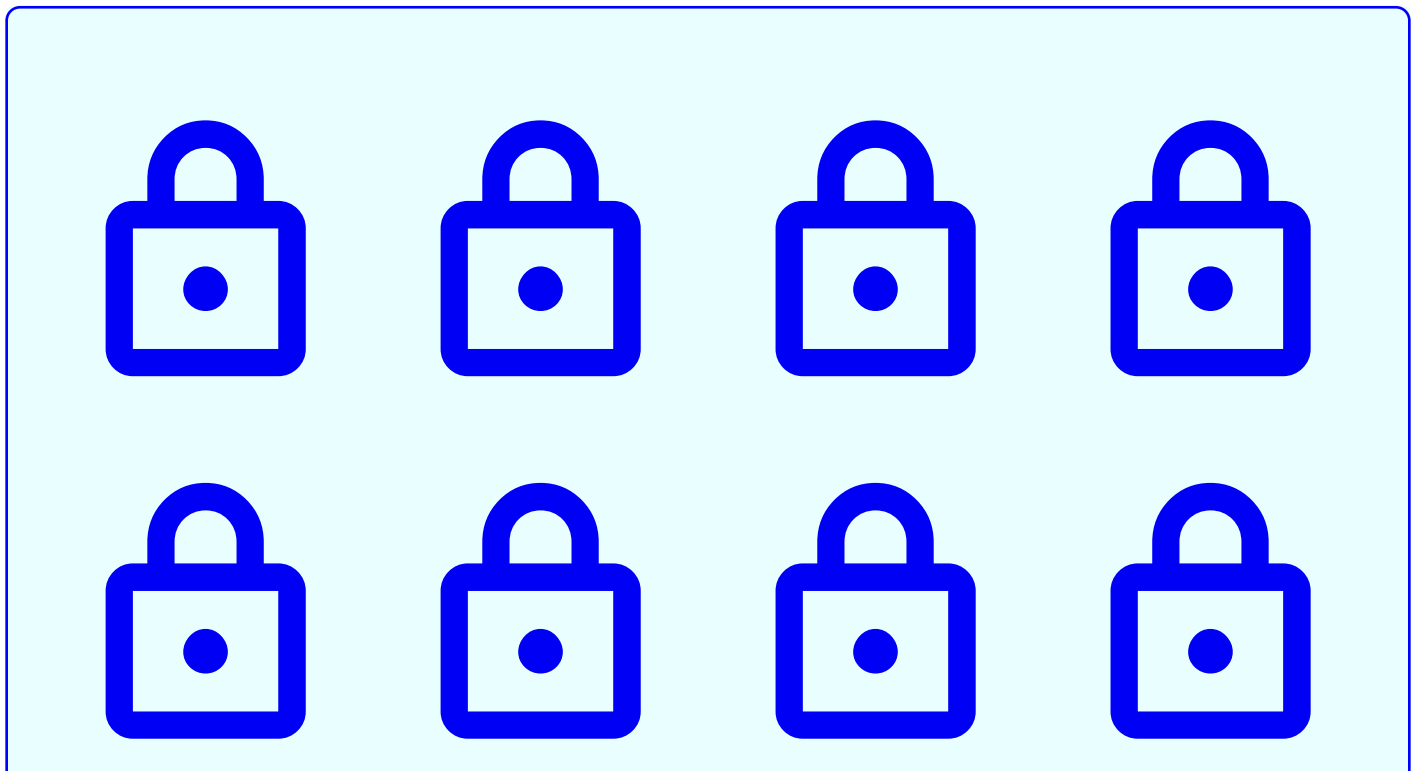
CONVOCATION



Safety by Design

A comprehensive methodology and exploration on design, policy and technology interventions to generate better user safety on platforms.

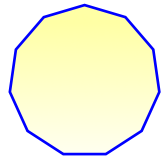
AUTHORS:
CAROLINE SINDERS, ANTONIA VALENCIA AND SAM SMITH / FEBRUARY 2026

RESEARCH + DESIGN

This report was commissioned by HateAid.

Table of Contents

Executive Summary



Social media harms are widespread and documented, especially harms related to digital hate speech, technology facilitated violence, online gender based violence, misinformation and disinformation and a myriad of other harms.

'Trust and safety', as a domain of expertise, serves a vital and important role in ensuring digital safety and combating these harms for platforms' users across technology. It's important to emphasize how 'trust and safety' is an area of knowledge contributed to over many years by activists, journalists, communities, policy experts, civil society, academia, and technologists.

Technology companies, while they employ teams labeled 'Trust and Safety' do not own the term or the expertise; it's an expertise shaped across many different areas of lived experience, knowledge, work and research. Within this, while trust and safety is a focus area of particular knowledge, those who work within the realm of 'trust and safety' (be it as a researcher or technologist), serve an important role in championing and furthering user safety. However, trust and safety teams at major social networks have consistently been an afterthought in terms of staffing, focus, and product building. This is further emphasized within the past few years, where, major platforms' trust and safety teams recently have faced dramatic cuts to their staffing size and funding[1], which in turn, impacts further impacts and decreases/degrades user safety online.

Digital violence is still an all too-common occurrence, with over 38% of women experiencing online violence, and 85% of women witnessing digital violence online, according to a report by the Economist Intelligence Unit[2]. This is why this report on Safety by Design has never been more urgent, and why it's necessary for interdisciplinary across policymaking, civil society, advocacy and academia to create foundational product elements, definitions, and standards for ensuring user safety across big tech and social media companies.

This research and report, conducted and created by Convocation Research + Design was commissioned by HateAid to develop comprehensive findings on the emergent space of Safety by Design, including a cohesive definition, and

1 https://www.nbcnews.com/tech/tech-news/big-tech-companies-reveal-trust-safety-cuts-disclosures-senate-judicia-rcna145435
2 https://onlineviolencewomen.eiu.com/

consolidating the myriad of important design, technology and policy based interventions, product augmentations and tooling to create a robust repository of safety design focused interventions that platforms should implement. At present, there is not a standardization of safety features for social media within academic, technology, policy and/or within the Trust and Safety world writ large; while most popular platforms have safety mechanisms such as muting, blocking, and reporting for harassment, not all do. Furthermore, additional safety mechanisms in product tooling, such as filtering of harmful comments, creating selective sharelists for content, turning off comments, etc are not universal across all apps, platforms or companies.

Users deserve consistency, agency, safety, and accountability online. Design alone will not fix the violence on social media but design serves an incredibly important role in allowing for user safety. New and updated design, policy, and regulation must work together in tandem to ensure safety for all users. It's time for technology to work for all users, not against them.

METHODOLOGY

This research employed a mixed-methods approach combining literature review, interviews with privacy and security experts, case study development, a validation workshop and a taxonomy utilizing data visualisation, and analysis of methods, interventions, design components and design patterns within the taxonomy as categories.

Across 2025, we conducted an extensive literature review, interviewed 7 experts across policy, technology, design, academia and human rights; and held an interactive workshop for feedback on our findings with 10 participants. This research report references 110 papers, articles, essays, and policy briefs[3] to create a cohesive and inclusive definition for 'Safety by Design' and source design patterns, design components, attributes and interventions, etc., to create a taxonomy of Safety by Design elements.

To illustrate the current pitfalls and lack of safety, security and privacy on social media platforms, and showcase how the taxonomy can help provide solutions, we created a series of six anonymized case studies that includes a detailed narrative of how harm unfolds across multiple platforms and specific examples of how platforms could implement Safety by Design by utilizing the taxonomy. These case studies and the Safety by Design Taxonomy are described in our larger report.

SAFETY BY DESIGN

While the term "Safety by Design" has entered the lexicon of discourse across academia, policy and regulation

3 Much of the content in this report comes from our extensive literature review, including our assembling of different design interventions and policy suggestions. This section's literature review is a broad overview of the entire literature assembled for this project.
All work cited in this report is listed in our bibliography section in the Annex.

in regards to digital platforms and harms, there are numerous definitions and alternatives of 'Safety by Design.' This definitional ambiguity can create confusion. In our qualitative interviews and workshops, different stakeholders had different definitions and opinions on Safety by Design and what it should entail. A series of principles guided the formation of our proposed definition, including: safety is proactive, not reactive; safety should center and build on security and privacy by design; it should center marginalized communities and "Design from the Margins" methodology; enable safety, not just prevent harm; it must be intersectional; it should focus on agency; center legibility and transparency; ensure active maintenance, research and response; and center accessibility.

Presented here is a proposed Safety by Design definition:

Safety by Design is an approach and methodology (including a related taxonomy) that centers safety for individual users, groups, collectives, and communities from the beginning of the technology, design and ideation process of software and hardware development — including and especially focused on very large online platforms. Safety by Design must ensure that even when groups or communities have conflicting needs, the most impacted group is still being centered; this is where working with civil society groups, community organizations, activists and design experts is key.
Safety by Design includes, but is not limited to, following accessibility design standards, privacy by design, security by design, and designing to follow compliance.
By embedding safety, agency, legibility, and ease of use into every product and safety feature, systems can be built that prioritize humanity from the start. Similar to security by design, Safety by Design is not an afterthought but must be engaged with before, during and after the product development process. Safety by Design means active maintenance, research, and response. As seen with security by design, this means that "during the initial planning, architecture, and design phases, teams proactively identify potential security threats and build protective measures directly into the fundamental structure of what they're creating."

For Safety by Design, this means that during the initial planning, architecture, and design phases, teams proactively identify potential harms, attack areas, and ways that the system can be weaponised through an intersectional lens (such as centering harms like online gender based violence (OGBV), technology facilitated violence (TFV), bullying, etc) and build protective and mitigation measures directly into the fundamental structure of what they're creating.

THE SAFETY BY DESIGN TAXONOMY

The Safety by Design Taxonomy has been developed as a **reference tool or library** to archive, analyse, classify, and understand the different interventions that can contribute to safer digital ecosystems to mitigate and prevent harm. This taxonomy has four broad categories with over 350 data points and examples within it. It draws from existing taxonomies[4], expert consultation, and an extensive literature review. The taxonomy exists to guide policymakers, researchers and designers to pre-existing knowledge and proposed interventions, patterns, principles and design products to use and reference in their work, to inspire new types of Safety by Design, and **should be used as a reference to hold platforms accountable for their lack of safety interventions.**

IMPLEMENTATION MATTERS FOR SAFETY BY DESIGN

No definition or accompanying taxonomy can be exhaustive in its design or knowledge repository, but this is where implementation is key in ensuring a true Safety by Design process. By situating Safety by Design in best practices from human rights, Design from the Margins, and engaging with principles like transparency and knowledge transfer; these best practices are needed in implementing a Safety by Design process that works. The implementation and (mis)interpretation of the Safety by Design definition came up in interviews and in our workshop as a concern and a potential reality for platforms trying to engage in a Safety by Design process. Workshop attendees suggested contextual information and principles to accompany the Safety by Design definition and taxonomy to ensure optimal implementation with suggestions including:

- Implementation matters
- Who's listened to in the design process impacts what is made and how
- Avoiding 'generalizing' communities, engage with real communities in a Design from the Margins style process, and recognize that not everyone in the same community will have the same exact needs
- Documentation of decision making is needed for ongoing maintenance of products
- Ensuring those with the least privilege are brought in and considered in the design process
- Understand that the Safety by Design taxonomy and definition itself must have ongoing maintenance; nothing is static and everything can and should be updated, must like the products that Safety by Design aims to guide

4 https://tfgbv.humane-intelligence.org/ NDI Landscape Tracker: TFGBV Interventions * tracker https://docs.google.com/spreadsheets/d/1SFJ-ZWryxYN9REXGCdM7dBjoBQM9-JHLJgC6osK0Kzo/

RECOMMENDATIONS

Safety by Design is an on-going, interactive process that requires community and expert input from activists, journalists, policymakers, civil society and academia, and for real transparency and collaboration from platforms to fully participate and engage in these Safety by Design suggestions. Many of the interventions, mitigations, and solutions documented in the Safety by Design taxonomy have been demanded by impacted individuals, communities and experts for years. Most, if not all of these suggestions have been threat modeled, analysed, and tested by outside experts and communities. To fully create more nuanced safety for users, it's time for platforms to truly implement comprehensive Safety by Design mechanisms that all users can access.

- Policy makers should use the definition, taxonomy and case studies together and pass stronger and specific regulation that centers actionable and concrete Safety by Design recommendations for platforms to implement now.
- Policymakers must not limit or target well known safety and security protocols, like encryption, in efforts to protect one community over another
- Many of the design interventions documented in the Safety by Design taxonomy have been demanded over and over again for years as types of design interventions at-risk users want. Platforms must provide accountability, transparency and information on what they've tried and why they are not implementing these suggestions. Furthermore, there must be ways for communities, activists and civil society to propose interventions even if platforms disagree.
- Power must be given to individuals to be able to enact agency and try to customize, and augment their digital experiences to their preferences.
- Demand that platforms invest more resources in trust and safety teams, and safety mitigations.
- Platforms must commit to truly engaging with Safety by Design. This means understanding that implementation matters, and so does a real commitment to the taxonomy, ongoing policy, and the definition. A platform cannot just try this work once, or treat the definition and taxonomy like a lightweight checklist to 'safety-wash' their work. Similarly, a platform should not center the needs of one, more privileged group over another, or only commit to tackling one type of problem and not tackle all of the different digital harms that communities face. Safety by Design does mean safety for all, not safety for some.

2. Introduction and Context

Social media harms are widespread and documented, especially harms related to digital hate speech, technology facilitated violence, online gender based violence, and the list goes on. Creating cohesive language, definitions, standards, and technology and design based interventions for safety that center all users has never been more paramount.
‘

Trust and safety’, as a domain of expertise, serves a vital and important role in ensuring digital safety and combating these harms for platforms’ users across technology. It’s important to emphasize how ‘trust and safety’ is an area of knowledge contributed to over many years by activists, journalists, communities, policy experts, civil society, academia, and technologists. Technology companies,

while they employ teams labeled 'Trust and Safety', do not own the term or the expertise; it's an expertise shaped across many different areas of lived experience, knowledge, work and research. Within this, while trust and safety is a focus area of particular knowledge, those who work within the realm of 'trust and safety' (be it as a researcher or technologist), serve an important role in championing and furthering user safety.

However, trust and safety at companies is rarely supported enough to tackle widespread harm at scale, and from the outside, seems to be a secondary or tertiary focus of platforms in general — alongside platforms gaining knowledge about how their systems work, but rarely sharing that knowledge outward to the trust and safety communities at large. Additionally, in the past few years, the Trust and Safety teams across major social networks have seen their budgets slashed and staff let go. What started when Elon Musk took over Twitter has spread to Meta, and Google products as well[5]. Trust and safety teams serve a vital and important role in ensuring digital safety for all users, and when teams and budgets are cut, user safety decreases and degrades.

Digital violence is still an all too-common occurrence, with over 38% of women experiencing online violence, and 85% of women witnessing digital violence online, according to a report by the Economist Intelligence Unit[6]. This is why this report on Safety by Design has never been more urgent, and why it's necessary for interdisciplinary across policymaking, civil society, advocacy and academia to create foundational product elements, definitions, and standards for ensuring user safety across big tech and social media companies.

This research and report, conducted and created by Convocation Research + Design was commissioned by HateAid to develop comprehensive findings on the emergent space of Safety by Design, including a cohesive definition, and consolidating the numerous important design, technology and policy based interventions, product augmentations and tooling to create a robust repository of Safety by Design based interventions that platforms should implement.

At present, there is not a standardization of safety features on social media; while most popular platforms have safety mechanisms such as muting, blocking, and reporting for harassment, not all do. Furthermore, additional safety mechanisms in product tooling, such as filtering of harmful comments, creating selective sharelists for content, turning off comments, etc are not universal across all apps, platforms or companies.

In similar ways in which all modern cars have safety features such as airbags and seatbelts, so too should social networks have 'seatbelts' of online safety or cohesive and comprehensive standards

5 https://www.nbcnews.com/tech/tech-news/big-tech-companies-reveal-trust-safety-cuts-disclosures-senate-judicia-rcna145435

6 https://onlineviolencewomen.eiu.com/

found across all companies. A robust Safety by Design definition and design requirements alongside policy standards to ensure comprehensive user safety across social media is needed. While design and policy alone cannot solve the problems of technology facilitated violence, design and digital products serve useful roles in providing spaces for users to quickly augment, mitigate and, in some cases, prevent harm that they are facing; these tools help quickly return agency to users.

Users deserve agency, safety, and accountability online. Design alone will not fix the violence on social media, only large structural, systemic and legal changes can combat those harms, but design, like seat belts, serves an incredibly important role in allowing for user safety. New and updated design, policy, and regulation must work together in tandem to ensure safety for all users, but especially marginalized and vulnerable users.

METHODOLOGY

This research employed a mixed-methods approach combining literature review, interviews with privacy and security experts, case study development, a validation workshop and a taxonomy utilizing data visualisation, and an analysis of methods, interventions, design components and design patterns within the taxonomy as categories. Each research method provided insights into building out this research project.

Across 2025, we conducted a literature review, interviewed 7 experts across policy, technology, design, academia and human rights; and held an interactive workshop for feedback on our findings with 10 participants. This research report pulls from an extensive literature review of 110 papers, articles, essays, and policy briefs[7] to create a cohesive and inclusive definition for 'Safety by Design' and source design patterns, design components, attributes and interventions, etc to create a taxonomy of Safety by Design elements.

To illustrate the current pitfalls and lack of safety, security and privacy on social media platforms, and then showcase how the taxonomy can help provide solutions, mitigations, and suggestions for improvements for these shortcomings, we created a series of six anonymized case studies that includes a detailed narrative of how harm unfolds across multiple platforms over time, specific examples of how platform features enable or fail to prevent harm, and suggestions of interventions pulled directly from the Safety by Design taxonomy that illustrate types of design components, design patterns, attributes or interventions that could address, mitigate or counteract the harms found specifically within each case study.

7 Much of the content in this report comes from our extensive literature review, including our assembling of different design interventions and policy suggestions. This section's literature review is a broad overview of the entire literature assembled for this project.
All work cited in this report is listed in our bibliography section in the Annex.

Expert Interviewees

Name	Affiliation	Area of Expertise
Hera Hussein	Executive Director, Chayn	Online gender based violence, and trauma informed design expert
Dr. Oliver Haimson	University of Michigan	Professor on digital harm
Theodora Skeandas	Formerly of Twitter's Policy Team	Policy and online gender based violence expert
Afsaneh Rigot	The De|Center, founder and principal researcher	Human rights, online gender based violence
Bojana Kostic	Co-founder, Pen to Paper	Online gender based violence and policy expert with a focus on EU and Western Balkans
Anonymous	Former Big Technology UX designer	Design and online gender based violence
Eva Galperin	Director of Cybersecurity, the Electronic Frontier Foundation	Security, Head of the EFF's Threat Lab and Stalkerware expert

Interactive workshop attendees

Name	Affiliation	Area of Expertise
Roja Heydarpour	The De|Center	Human rights, design
Eriol Fox	Open Source Design	Design
Cala Del Rio	Fora Design	Design
Łukasz Krol	International Committee of the Red Cross	Security
Stephanie Mikkelson	UNFPA	Policy, online gender based violence
Dyan Cortez	Open Technology Fund	Security, design
Luisa Ortiz Perez	Vita Activa	Online gender based violence, policy
Sydette Harry	Feminist and harassment expert	Online gender based violence expert
Kayee Au	Kumqat Design	Design
Brittan Heller	Stanford University	Policy, online gender based violence

3. Literature and Landscape Review

This research project started with a multipronged goal: to create a nuanced and comprehensive Safety by Design definition, and a taxonomy of design requirements, components, design patterns, current products that can increase safety, security and privacy, and products to be built to create safety, security and privacy. By engaging in a comprehensive landscape and literature review, our goal was to collect many of the ideas, theories, and suggestions across a 15 year time span, in part to illustrate the robustness of ideas for user safety that have been produced, including ideas that are currently being implemented by platforms (including non-social media platforms), and to also illustrate the gap of safety mechanisms on social media. It's important to emphasize that so much good work on user safety in regards to technology facilitated violence and

online gender based violence has been conducted, and it's equally important to acknowledge that those important ideas are not being engaged with enough by social networks.

By creating this comprehensive Safety by Design definition and taxonomy, our goal is to utilize these findings to hold platforms accountable by advocating for actionable steps, and improvements on design, technology and policy that these platforms should take.

SAFETY BY DESIGN: ITS BACKGROUND AND AN ANALYSIS

Safety by Design, as a concept, has emerged from multiple regulatory and academic traditions, including Australia's eSafety Commissioner pioneering a formal Safety by Design standards starting in 2018. From the very beginning, Australia's Safety by Design standards states "Safety by Design puts user safety and rights at the centre of the design and development of online products and services[8],"arguing that design and development of online products have a clear and specific role to play in creating user safety.

While countries like Australia are determining their own Safety by Design definitions, Safety by Design is also an emergent design, technology and policy theory and builds on pre-existing methodologies and theories, like Privacy by Design which was developed by Ann Caukvovian in 2006 for the Information and Privacy Commissioner of Canada[9].

In order to build a comprehensive definition of Safety by Design, this paper looked at numerous design theories across security, privacy, ethical design[10] [11], design justice[12], responsible design[13] and many others. Even if not explicitly called 'Safety by Design', principles related to transparency, privacy, user agency, human rights, and feminist design are elements that center safety, accountability and agency for users, even if the names of these design theories lack the explicit term 'safety.' The Orbits Field Guide by Chayn is an example of this, and outlines an expanded framework encompassing eight core principles: "Safety, Agency, Equity, Privacy, Accountability, Plurality, Power Redistribution, Hope" that expand Safety by Design beyond technical specifications to encompass trauma-informed, survivor-centered design[14].

Other papers have also looked at design theories that are building blocks of 'Safety by Design' but that do not use the word safety explicitly in its naming or principles. The paper, Using 'Safety by Design' to Address Online Harms[15] by John Perrino, outlines how Safety by Design as a theory builds

8 https://www.esafety.gov.au/industry/safety-by-design

9 https://iapp.org/media/pdf/resource_center/pbd_implement_7found_principles.pdf

10 https://www.artefactgroup.com/case-studies/ethical-explorer-pack/

11 https://99designs.com/blog/tips/ethical-design/

12 https://designjustice.org/

13 https://www.mozillafoundation.org/en/blog/how-create-stronger-safer-online-communities/

14 https://chayn.notion.site/Orbits-a-global-field-guide-to-advance-intersectional-survivor-centred-and-trauma-informed-interv-8d8dc6a1436543b8b25af63ddafaa409

upon the concepts 'Privacy by Design[16]' and 'Security by Design', which are cybersecurity principles[17] but do not necessarily center 'safety' in their descriptions. For example, 'security by design' represents a paradigm shift from post-incident response, meaning after an incident has occurred, to proactive risk mitigation embedded in system architecture, which is related to safety mitigation even if not explicitly mentioning 'safety'. Going further, this same paper describes how Safety by Design "offers a more proactive approach for policymakers to address ever-evolving online safety issues" fulling from the aforementioned Security by Design principles. To achieve Safety by Design, Perrino argues that principles like platform accountability, user empowerment and transparency must be core parts of the theory.

For example, platform accountability requires "risk assessments, independent audits, and transparency reporting[18]". User Empowerment provides features offering "control over who can contact them," "ability to filter harmful content," and "options to opt out of algorithmic recommendations". Transparency and research access mandates "that platforms share data with vetted researchers".

Other design frameworks were surfaced that explicitly described safety elements but focused more on human rights and decolonization, such as "Design from the Margins' by the De|Center and its founder and principal researcher Afsaneh Rigot, the Human Rights Centered Design theory by Caroline Sinders and Natalie Cadranel[19], Design Justice[20], and many others. In one of the foundational papers of Design from the Margins, Rigot describes how products need to be tested to be safe for all users, including the most marginalized and vulnerable whose harms should not be considered edge cases.

Safety by Design means following this ethos laid out by Design from the Margins in which "we expect architects to design buildings and bridges to withstand gale force winds and heavy loads, so too should we expect the companies whose products are the vehicle for free expression and access to information to design sensitive, resilient technologies for the benefit of all, whether they face government repression and social stigma or not[21]."

Human Rights Centered Design, a curriculum and design theory, also describes specific principles found in other Safety by Design frameworks, such as security, safety, and accessibility.

15 https://www.brookings.edu/articles/using-safety-by-design-to-address-online-harms/
16 https://iapp.org/media/pdf/resource_center/pbd_implement_7found_principles.pdf
17 https://www.brookings.edu/articles/using-safety-by-design-to-address-online-harms/
18 https://www.brookings.edu/articles/using-safety-by-design-to-address-online-harms/

19 https://www.humanrightscentered.design/
20 https://mitpress.mit.edu/9780262043458/design-justice/
21 https://www.belfercenter.org/publication/design-margins

The Human Rights Centered Design Principles[22]:

- **Integrates human rights directly** into the lifecycle of R&D (research and development)
- Centers the **needs of communities**
- **Security-focused:** ensuring the protection of users is key, recognizing that security and safety are fundamental human rights
- **Accessibility:** the tools should be localized and context-specific so they are usable and accessible to the communities they serve
- **Innovative workflow** that is responsive and adaptable to communities' needs
- Engages in **cooperative and collaborative co-design** (moves beyond participatory design)
- Appreciates **plurality of knowledge creation** and understands that the expert is an individual, a community, a coalition, an activist, a cooperative, a researcher, and an academic, e.g. the expertise is accessible to everyone, and it exceeds the boundaries of traditional institutions that purport to wield special authority over knowledge production
- **Public / community intelligibility:** it makes information broadly intelligible by avoiding jargon and unnecessary complexity

22 https://www.humanrightscentered.design/introduction/intro

Design Requirements, Design Failures and Utilizing Design to Mitigate Harms

Many of the papers we analysed and read in our literature and landscape review directly mentioned, audited, critiqued, and proposed new forms of design products and design interventions. All of the design examples, design patterns, and design components are listed, and organized in our taxonomy, which is further described in section 4b, and provides an indepth breakdown into all of the different themes, and design patterns surfaced in this report.

However, this current section exists to highlight the importance of design in safety, security and privacy. As mentioned earlier, design alone cannot solve technology facilitated violence, but design combined with technology and policy that is constantly evolving, and updating based on new findings, emergent threats, and needs has a particular and key role to play in user safety. From our expert interviews, workshops and literature review, what emerged was a holistic understanding that design has an important role to play but that safety cannot be designed or 'solved' alone in isolation, it requires design to be wrestling with constant fundamental tensions of inequity, challenging existing power structures present in society and technology and continuously adapting to evolving harms to better serve users.

IMPLEMENTATION MATTERS

The insights from interviewed experts revealed that Safety by Design is both more urgent and complex than initially apparent, and highlighted that design and how it's implemented and engaged with matters. While there was a broad consensus on core principles, implementation of these principles into platforms' design and policies is harder and requires strenuous and dedicated work from the platforms. Yet, experts also provided some examples where design choices make meaningful differences demonstrating that Safety by Design is a practical framework that can guide specific and actionable improvements. Most importantly, experts emphasize that Safety by Design is not a one-time thing, it's an ongoing process.

Afsaneh Rigot unpacked some of the difficulty in implementing design for safety. She explained, "so for us from a decentered design view, as a framework when doing implementation of tech changes and design from the margins, you need to have a method to translate the needs and wants, the harms or the documentation directly from communities to implementation. And sometimes it's not easy and you need folks who are skilled [in that implementation]." This translation and implementation work Rigot is describing requires people who can bridge community expertise, design and technical capabilities while also working with individuals and then experts who understand both the lived realities of marginalized users and the constraints and possibilities of platform architecture. However, this is where it's key to understand that design is one of the interdisciplinary domains that can allow for community expertise, technology, policy and product design to all work together to implement these exact changes with the types of experts Rigot is describing.

UNDERSTANDING THE POWER OF DESIGN AND LEVERAGING IT TO SHAPE OUTCOMES

Theodora Skeandas noted how important design is, especially in shaping interactions on platforms, and how design, as a medium, holds a special power in facilitating safety. Skeandas said, "If you look at what I call curation, it really means customization. It includes design. That's 25% of all recommendations by civil society to tech companies on ways to address the tfgbv [technology facilitate gender based violence] harms are all about design, right? Designing the systems so that people can curate, customize, better experiences for themselves." Allowing for user agency, and making all types of design led changes, from the micro to the macro, have real impacts on user safety and user experience.

Even small design choices, like nudges, can have a widespread impact. Skeandas spoke on the effectiveness of nudges: "Nudges are very effective, right? 30% effective, typically. So like, more nudges? Like, 'is this something you really wanted to say?' 30% of people will not say it if they get a nudge." This suggests that relatively simple design interventions,

such as nudges and text prompts to encourage users to reconsider before posting potentially harmful content can meaningfully reduce harm.

OTHER EXAMPLES OF DESIGN INTERVENTIONS TO CREATE SAFETY

Some of the earliest papers in our bibliography date back to 2015[23], and feature similar small but concrete design tooling and product component suggestions, like nudging, to create safety on apps like Facebook and Twitter. Some of these suggestions have been implemented such as allowing screenshots to be submitted as evidence, but others like allowing users to 'batch report' or group multiple types of harassment into one report, and have more nuanced privacy settings, still have not been engaged with.

Across the many papers reviewed for this report, many different types of design suggestions appeared, with **micro suggestions for improving pre-existing tools** such as including one small tweak here or there, to **more medium and macro suggestions** on pre-existing tools that illustrate a nearly complete redesign. For example, Pen America's "Shouting into the Void[24]" offered suggestions on how to improve current reporting dashboards, and even proposed new dashboard designs and functionality that any platform could implement.

Some recently published papers emphasized **new technology innovations** like better developed quality filters and leveraging AI automation, but also made sure to reference **safety interventions that have long been demanded by activists**, such as "opt-out rather than opt-in for privacy settings." Opt-out privacy settings means, in design speak, that the maximal privacy or the most privacy focused settings are set as a default, and if a user wants less privacy, they need to go and change the settings. Presently, most platforms set privacy settings as opt-in, with the default settings being the most public option (e.g. the least privacy

- Micro-suggestions to improving pre-existing tools
- Medium and macro suggestions to pre-existing tools
- Leveraging new technology innovations
- Uplifting historical safety requests that are still not implemented
- New products for safety
- Have platform design teams engage with principles, gaps and needs highlighted from researchers to develop into products

23 https://www.washingtonpost.com/news/the-intersect/wp/2015/05/14/the-conclusive-expert-guide-to-saving-twitter-from-its-trolls/

focused option). One paper suggests that when users choose a less private option, using that moment as a form of education and "prompt[ing users] to consider the trade-offs before confirming their selection. For example, platforms can follow this recommendation by requiring that location sharing on maps should be turned off by default[25]."

Additionally, other papers suggested **nuanced and new types of design products** and tooling that could exist such as 'safety nodes' or one click nuanced privacy settings to protect users , alongside calls **of larger principles, gaps or needs that product teams should explore and build into products.** An example of these larger principles are calls for better and more quickly identifying abusive material. While that is not a product in and of itself, it's a principle that a designed product could follow and it's a need that users face in which their harassment is not recognized quickly enough or it's not recognized at all. Another Pen America report "No Excuse for Abuse[26]" encourages platforms to redesign to "empower users rather than just punish abusers", which is also a principle. That same paper does offer a variety of specific design suggestions and principles including

- A "shield and dashboard" system that quarantines abusive content for safe review
- "safety modes" for one-click privacy protection
- "rapid response teams" of trusted allies to help manage abuse
- automatic documentation tools for legal evidence
- transparent penalty systems with escalating consequences
- trauma-informed reporting for coordinated attacks.

Each suggestion then includes in-depth ideation on how this could be achieved via design. For example, 'safety nodes' includes a thematic section around the principle of providing users with "robust, intuitive, user-friendly tools to control their privacy and security settings." "Safety nodes" are then described to be: multiple, distinct configurations of privacy and security settings that they can then quickly activate with one click when needed.

"No Excuse for Abuse" goes further to illustrate that suggestion and place it in reality as a call to action by writing "Twitter and Instagram should give users the option to fine-tune existing safety modes ("protect my tweets" and "private account," respectively) after users activate them. These modes are currently limited in functionality because they are binary (i.e., the account is either private or not). Facebook should introduce a safety mode that allows users to go

24 https://pen.org/report/shouting-into-the-void/
25 https://cdn.prod.website-files.com/68ba1281b6406dbd18597dab/69209b4a3bbf263928503786_Prevention-by-Design-A-Roadmap-for-Tackling-TFGBV-at-the-Source.pdf
26 https://pen.org/report/no-excuse-for-abuse/

private with just one click, as Twitter and Instagram have already done, while also ensuring that users can then fine-tune specific settings in the new safety mode." This is just one of many examples from the paper that situates a specific user need, and then grounds that analysis in how practically and specifically a company can meet that need via their design.

MORE EXAMPLES FROM THE FIELD OF IMPROVING PRE-EXISTING PRODUCTS AND PLATFORMS

Similarly, the work by the De|Center, who are the creators of the Design from the Margins design theory, often works with pre-existing products and companies to critique and remake products to better protect 'marginalized' or 'decentered users[27].' For example, their paper "Design from the Margins[28]" describes how Afsaneh Rigot, the founder and principal researcher of the De|Center, worked with human rights organizations and activists and MENA LGBTQ communities and created design suggestions to better serve and protect those communities.

They created this anonymized scenario that came out of their research to guide design research:

> "A queer Syrian refugee in Lebanon is stopped at a police or army check point for papers. They have their phone arbitrarily searched. The icon of a queer app, Grindr, is seen. This outs the individual as queer in the officer's eyes. Other parts of the individual's phone are also then checked, revealing what is deemed as "queer content". The individual is taken in for further interrogation and subjected to verbal and physical abuse. They now face sentencing under Article 534 of the Lebanese Penal Code and face potential imprisonment, fines and/or revocation of their immigration status in Lebanon."

From this anonymized scenario, community input and input of design and security professionals, created a series of specific design suggestions for Grindr, which brought about real world changes like the 'discreet icon app' and the 'self destruct PIN'.

27 "Decentered users are the groups most at risk and under-supported in the relevant context," https://www.de-center.net/what-is-design-from-the-margins1-1-1

28 https://www.belfercenter.org/sites/default/files/pantheon_files/files/publication/TAPP-Afsaneh_Design%20From%20the%20Margins_Final_220514.pdf

Other design ideas this project came up with were:

- best practice outlined on mobile application security and API security
- more creativity and privacy respecting user verification models
- unsend and disappearing messages to help ensure better secure communications especially against entrapping police fake profiles, and avoid backlogs of chat and photo data if devices are under inspection as well
- blocking or providing alerts when another user takes a screenshot as police rely heavily on screenshots when using fake profiles to entrap queer individuals
- providing more up-to-date, geolocation-based safety guides (instead of generic tips) that reflect country-specific laws, practices, events, with consistent feedback from local organizations;
- direct lines of communication in cases of arrest or emergencies via apps

PRINCIPLES AND REQUIREMENTS TO INFORM DESIGNERS AND DESIGNING PRODUCTS

Across the papers, workshops and interviews, experts highlighted different types of principles to guide design, and that also centered users' specific needs, particularly users who are actively being targeted or underthreat. These principles appeared as major, recurring themes across our research, but for a comprehensive look at all principles and design theories surfaced in this research, please reference our Safety by Design Taxonomy in **Section 4b.**

The principles mentioned in this section can be translated into specific design takeaways, but they exist more so as design and technology principles to guide platforms and products to center Safety by Design. All of these principles must be adequately and deeply engaged across the entirety of the platform and its related products to fully embody a Safety by Design ethos. Only engaging in one of these principles for platform design, or only utilizing these principles for one part of the platform is not engaging with Safety by Design and it would not then fully address, mitigate or prevent harm within the platform.

Principles include:

- That design must be usable, accessible, and legible, and easy to find
- That design and technology be trauma informed
- Centering opt out over opt in in terms of safety, privacy and security settings to ensure the most safe user experience
- Design should take into account the emotional state of the user
- Design must repeatedly center the most marginalized or de-centered user
- Design should actively threat-model and understand and prepare for all different types of nuanced harm

4. Safety by Design: The Definition and Taxonomy

4a. Creating a More Cohesive Safety by Design Definition

While the term "Safety by Design" has entered the lexicon of discourse across academia, policy and regulation in regards to digital platforms and harms, there are numerous definitions and alternatives of 'Safety by Design.' This definitional ambiguity can create confusion. In our qualitative interviews and workshops, different stakeholders had different definitions and opinions on Safety by Design and what it should entail. This section presents a proposed definition of Safety by Design, unpacks the principles and frameworks that informed it and illustrates why a more cohesive definition matters for advocating for user safety marked within real measures and interventions.

Our proposed definition:

Safety by Design is an approach and methodology, including a related taxonomy, that centers safety for individual users, and also groups, collectives, and communities from the beginning of the technology, design and ideation process of software and hardware development, including and especially focused on very large online platforms.

Safety by Design must ensure that even when groups or communities have conflicting needs, the most impacted group is still being centered; this is where working with civil society groups, community organizations, activists and design experts are key. Safety by Design includes, but is not limited to, following accessibility design standards, privacy by design, security by design, and designing to follow compliance.

By embedding safety, agency, legibility, and ease of use into every product and safety feature, systems can be built that prioritize humanity from the start. Similar to security by design, Safety by Design is not an afterthought but must be engaged with before, during and after the product development process. Safety by Design means active maintenance, research, and response. As seen with security by design, this means "that during the initial planning, architecture, and design phases, teams proactively identify potential security threats and build protective measures directly into the fundamental structure of what they're creating."

For Safety by Design, this means that during the initial planning, architecture, and design phases, teams proactively identify potential harms, attack areas, and ways that the system can be weaponised through an intersectional lens (such as centering harms like online gender based violence (OGBV), technology facilitated violence (TFV), bullying, etc) and build protective and mitigation measures directly into the fundamental structure of what they're creating.

IMPLEMENTATION, DOCUMENTATION, AND UNDERSTANDING THE NUANCES OF COMMUNITY MATTERS:

It's imperative to note that no definition can be exhaustive, and all encompassing in its scope, but this is where implementation is a key part of ensuring a true Safety by Design process. By situating Safety by Design in best practices from human rights, Design from the Margins, and engaging with principles like transparency and knowledge transfer; these best practices are needed in implementing a Safety by Design process that works.

The implementation and (mis) interpretation of the Safety by Design definition came up in interviews and in our workshop as a concern and a potential reality for platforms trying to engage in a Safety by Design process. One attendee of our workshop, Stephanie Mikkelson, a tech-facilitated gender based violence programme specialist with the UNFPA, pointed that this also includes knowing "when to not build if the risk is too high."

Other attendees like feminist thinker and harassment expert Sydette Harry pointed out power differentials inside of companies themselves between the company, its employees, and the victims impacted, and documentation of harms. Sydette explained, "The idea of exactly how power relationships function and how those don't make it into the design briefs. We often describe stakeholders as having equal power in design briefs because we like them, they're pretty. But we do not have the point [of view] of the rural farmer [who] does not have as much power as the technological person at the city center, that the person who does not exist along the gender binary does not have as much power as the CEO. And [those power differentials] becomes undocumented."

IMPLEMENTATION MATTERS

- Who's listened to into the design process impacts what is made and how
- Not all communities are the same, and not all in communities have the same needs
- Documentation of decision making is necessary for ongoing maintenance
- Ensuring those with the least privilege are brought in and considered in the design process
- Avoiding 'generalized' communities and engage with real communities in a Design from the Margins style process
- Understand that the Safety by Design taxonomy and definition must have ongoing maintenance; nothing is static and everything can and should be updated, must like the products that Safety by Design aims to guide

Often, what happens, as Sydette describes, is that the power of choices inside of the company, and of who's values or pain was centered and whose was not impacts what is made, but that is often not documented. This documentation is necessary for design and for safety. Sydette elaborated, "Documentation is often how we look at who is responsible and who is accountable. So it's not just the design, it's the does everybody know why that design is here? Does we have the power and who can go back and get it as well?"

Another attendee, Roja Heydarpour, head of communications the DeICenter, brought up power dynamics within communities, building upon the conversations around privilege, and access by avoiding the notion of generalized communities, and ensuring that a lot of time and care are put into working with specific impacted communities to engage in a Design from the Margins process. Roja explained, "It takes a lot of time to be very clear and do the research and exactly document who these groups are and how to identify them, rather than just kind of generally speaking. And then you ask them also what kind of design changes are helpful for them. So that [changes] comes from the inside out, not from the outside in."

Another aspect of implementation and maintenance came from Sydette, which is clarifying that "Safety by Design" can understand dynamics and understanding that things change; in essence, no process, definition or document can be static.

Foundational Principles: What informed our Definition

SAFETY AS PROACTIVE, NOT REACTIVE

One of the fundamental principles underlying this proposed definition is that safety must be embedded from the beginning of the design process, not as an afterthought or retrofitted after harms have already occurred, and that maintenance of safety has to be ongoing throughout the product lifecycle, post the product launch, and for however long that product is used by users. Meaning, however long this product 'lives', the product must be maintained and updated to ensure safety; updates include not just code or policy updates, but also when new harms are discovered, the product must update to mitigate and address those harms. This emerged as a nearly unanimous perspective across expert interviews and in the workshop, and from the writers' own past experiences in the field. This principle is echoed in similar structure or language in Security by Design and Privacy by Design methodologies as well.

One of our interviewees, Dr. Oliver Haimson from the University of Michigan, articulated this clearly: "When I hear the term [Safety by Design], it makes me think that safety would be part of the initial design process rather than something tacked on at the end or, you know, thought about as separate from the rest of the design of the tool." This proactive approach requires a fundamental shift in

how platforms conceptualize, design and also present this to your users.

Eva Galperin, Director of Cybersecurity at the Electronic Frontier Foundation, emphasized this same point by stating that, "Safety by Design is not an afterthought. You have to do this from the very beginning. You need to center your most marginalized communities." This perspective, similar to Design from the Margins and Human Rights Centered Design, reframes safety not as a compliance checkbox or a post-launch feature, but as a foundational principle that shapes every decision in the product development lifecycle. This principle extends beyond initial product development and includes ongoing maintenance, research and response as well.

This proactive approach stands in contrast to how historically many technology platforms have addressed safety. For example, Twitter's delayed implementation of anti-harassment controls provides a cautionary example. As Hera Hussein, Executive Director of Chayn, explained, "Twitter could have put really strong anti harassment, anti bullying controls in there from the beginning. And they let it get so bad that when, you know, black actors started leaving the platform due to hate speech, that is when they brought it [those safety controls]."

Waiting until harms reach crisis levels before implementing protections is not how to enact Safety by Design. Once the point of harmful actions and behaviours are occurring and then impacting users at a wide scale, it becomes even more difficult to address those harms by policy and design, because those mitigations have to work at a large scale for those impacted which can be difficult to build and implement once a product has a large user base, old code and/or a complicated code base. True Safety by Design means anticipating how systems can be weaponized and building protections before the first user ever signs up, and responding to harms in real time to more easily address and mitigate those harms at scale.

BEYOND SECURITY AND PRIVACY BY DESIGN

This proposed definition builds on Security and Privacy by Design methodologies, similar to Australia's Safety by Design definition and Safety by Design principles proposed by John Perrino, but also includes comprehensive requirements as explained accompanying taxonomy. These frameworks have demonstrated value and response by being proactive, principles for user protection and technically strong.

Eva Galperin articulated the importance of security by design, but also needing more nuance, and more specificity. Galperin explained, "What you're doing is you're building on security by design, and I think security by design is great, but it is insufficient. What we need is actually something somewhat broader than that that also incorporates the need for privacy, the need for anonymity, the need for accountability, sort of also as a balance, and also something which centers the needs of marginalized people."

For example, a platform could have excellent encryption (which falls under security) and strong data minimization/ retention tooling (which could be tooling that allows users to turn off location tracking, for example, and that would fall under privacy) but that product could still be unsafe for users if it allows for coordinated harassment campaigns to drive off users from the platform; amplifies hate speech, or disinformation through recommendation algorithms; makes it difficult to block abusers or control who can contact you; provides no support/resources when non consensual intimate images are shared; and/or fails to address doxxing, stalking or image based abuse among other examples.

Safety by Design must address **all** of these harms, and require interventions including and beyond security and privacy measures. Safety by Design requires understanding how platforms shape and host social interactions, how power manifests in digital spaces and how design choices enable or prevent violence.

CENTERING MARGINALIZED COMMUNITIES: "DESIGN FROM THE MARGINS"

One of the most consistent themes that emerged across interviews was that Safety by Design must **center the experiences of those most vulnerable to harm.** This "Design from the Margins" approach fundamentally reorients who platforms prioritize in their design processes. As Hera Hussein articulates,

"Safety by Design, when I hear it, I think about a design process or design framework, design approach that really values the how the intentionally values and designs around the experiences of people that might be marginalized in society based on their identity … where their experiences are treated less like they might be outliers and more like that they are as important as the experience of someone from a majority like group that might have less of those experiences."

Afsaneh Rigot explained the ethical reasoning behind this approach, "that [it] really does require a moment of thinking about who you're prioritizing and it becomes kind of like an ethical framing of if you do actually prioritize these specific folks who have the heaviest harms of risks and frameworks and institutions, the heaviest, say policing, law enforcement, states, corporate actors and the least protection and you have created something safe for them, you can count your bets that it will also be something that's safe for the folks who are a little bit more protected."

Relatedly, Hera Hussein summarized this as, "safety for one group makes it safe for all" when platforms center the most marginalized. More specifically, when platforms design for their most vulnerable users or with highest risk such as women, LGBTQIA+ people, people of color, activists, journalists in hostile environments, sex workers, people fleeing intimate partner violence, and those facing the highest risks of harassment, surveillance, doxxing, or violence, they create systems that work better for *everyone*.

Theodora Skeandas, formerly of Twitter's Policy Team explained this design decision, "Prioritizing user safety, user rights, when thinking about design choices. So that when products are developed from the onset, they're developed with, you know, the user's best interests at heart. To me, it also speaks to the need to build inclusive spaces. On the issue of TFGBV, this would mean inclusive spaces for women and in particular women of color, women from other marginalized backgrounds, because those are the people who are often most at risk."

This principle challenges the dominant assumption in tech that designing for the "average user" is the most efficient approach. In reality, average users are imagined to be typically Western, male, English-speaking, able-bodied, and financially secure and design a system that works only for those users creates a system that systematically excludes and endangers those who don't fit that narrow profile. Design from the Margins inverts this logic, making those typically treated as edge cases the center of the design process, which in turn, makes the design stronger for all users.

ENABLING SAFETY, NOT JUST PREVENTING HARM

Another thematic area that emerged is how safety itself is perceived and framed. Hera Hussein explained, "With Safety by Design, I see it, I think in the way things have developed. People see Safety by Design almost like it's a form of punishment or like a very like violent form of accountability" From what Hussein is describing, it's that the perception that safety comes with punitive measures, and this misunderstanding might be a risk on how safety features are accepted within the broader audience. This approach of 'safety' with punishment, something that might be misinterpreted from language, marketing and posts from social media companies, might make 'safety' seem as restrictive rather than liberating.

Several interviewees emphasized reframing safety from a negative to a positive stance. Rather than focusing solely on stopping harmful experiences, Safety by Design should articulate what becomes possible when harm is mitigated, as Hera Hussein articulated, "I feel like what's important is that when we talk about safety, we don't just talk about the harmful experiences, but we talk about what we enable by stopping that harmful experiences. So it's like reframing that from stopping bad things from happening to letting good things happen, and making space for the good."

This reframing is critical because it shifts the conversation from restriction to empowerment, from limiting what users can do to expanding what they can safely accomplish. It acknowledges that safety is not the absence of risk, but the presence of conditions that allow people to express themselves, connect with others, and participate in digital spaces with agency. When platforms position safety features as empowering, calling for participation, community connection and care, users may engage with them more openly rather than if they are presented as constraints.

INTERSECTIONAL

This proposed definition explicitly references intersectionality, another theme brought explicitly and through inference in our interviews, workshops, and in our literature review. Harms online are not distributed randomly, and do not impact all users equally. They cluster at the intersections of marginalized and intersecting identities. Without an intersectional lens, Safety by Design interventions risk addressing harms facing relatively privileged members of marginalized groups while failing those at the margins of the margins.

AGENCY

One of the central pillars of the definition refers to agency, understood as the capacity for users to make meaningful choices about their safety, privacy and participation, also echoed through our interviews and literature review. Agency can be understood as the ability for users to make informed decisions, have more control over their information, interactions, and decision making online, including, knowing and controlling how their data is collected, and making nuanced decisions about who they interact with, who can see them, and who can interact with their content, and posts writ large. It also allows for users to navigate their experiences and configure their settings in a comprehensive manner, proactively and reactively, as well as having resources for when harm occurs, ways to seek help, appeal and access support.

However, there is a balance to consider. As several experts noted, platforms cannot simply give users infinite granular controls and consider their job done as that can overload or confuse users. As Dr. Haimson explained "that's the big paradox here is that the more control you give people, then the more complicated it is." The solution is not to abandon user control but to design systems where safe defaults protect users while providing comprehensible ways to customize for those who need it by focusing on usability and accessibility.

LEGIBILITY AND TRANSPARENCY

This proposed definition includes "legibility" as a core value. Legibility is the idea that users should be able to understand how systems work, what's happening to their data, and why decisions are made. This concept of 'legibility' is closely related to but distinct from transparency. Transparency means providing robust and accurate information. Legibility means providing information in ways that people can actually understand and act upon.

This dimension of Safety by Design addresses one of the most frustrating aspects of current platforms that emerged from our literature review and interviews: that users often have no idea what's happening to their content, their data, or their reports and why, with platforms operating like an impenetrable black of data. Legibility means designing systems where users can understand who can see their information and content, what actions have been taken, why content was removed, what data is collected and how it is used among other things.

ACTIVE MAINTENANCE, RESEARCH AND RESPONSE

Our definition also states that Safety by Design "means active maintenance, research, and response", meaning it's not a one-time implementation but a constant commitment to safety; this theme manifested from our interviews, and explicitly from Security by Design methodologies. As Bojana Kostic noted, "it's not a one off thing. I think it's a process, it's a continual review and that actually includes this kind of a loop of those who, individuals who are actually using the technology to kind of understand more their interaction with it."

This ongoing nature will reflect and evolve according to certain patterns of harm, for example:

- Threats evolve: Attack vectors that didn't exist when the platform was launched emerge over time. Each new vector requires new responses. A platform that implemented strong safety measures at launch but hasn't updated them in years is no longer practicing Safety by Design.
- Usage patterns change: How people use platforms shifts over time, creating new risks. Features designed for one purpose could evolve over time and get repurposed in ways that could enable harm. We need to respond to these changes.
- Communities provide feedback: Users experiencing harm know things platform teams might not. Safety by Design requires listening to this feedback and responding, not treating initial implementations as final.
- Context matters - Active research means continuously evaluating whether safety measures serve diverse populations
- Regulation changes

ACCESSIBILITY

Our definition explicitly includes "following accessibility design standards" as a component of Safety by Design. Accessibility is fundamental to safety. As Afsaneh Rigot stated in our interview, "if it's inaccessible to people that need it most, is it safe?" This reframes accessibility from a separate concern to a core safety question. A panic button that only works for sighted users is not safe. A reporting system that requires complex navigation is not safe for people with cognitive disabilities, those new to technology or those new to the platform. Relatedly, Privacy settings that assume English fluency are not safe for non-English speakers.

4b. Safety by Design Taxonomy

The Safety by Design Taxonomy has been developed as a structured tool to analyse, classify, and understand the different interventions that can contribute to safer digital ecosystems to mitigate and prevent harm. This taxonomy draws from existing taxonomies[29] expert consultation and an extensive literature review, but it is considered emergent and 'incomplete' as new interventions, principles and mitigations for Safety by Design appear every day. The Safety by Design Taxonomy aims to be a reference tool or library to guide policymakers, researchers and designers to pre-existing knowledge and proposed interventions, patterns, principles and design products to use and reference in their work, to inspire new types of Safety by Design, and be used as a reference to hold platforms accountable for their lack of safety interventions. This taxonomy exists as a data visualization of hundreds of types of Safety by Design principles, products, technology and tooling platforms could use, and that have been proposed by experiments and communities for many years.

FOUNDATIONS AND INFLUENCES

Rather than creating an entirely new framework, the Safety by Design Taxonomy synthesizes and structures existing knowledge from a series of sources like the NDI Landscape Tracker and TFGVB Humane Intelligence Taxonomy and information from an extensive literature review. The goal is to provide an organizing framework and structure that collects and archives all of these different knowledge sources. As mentioned in Section 3 - Literature and Landscape Review, the key areas of reference include:

- Tech-Facilitated Gender-Based Violence (TFGBV) Frameworks
- Design from the Margins
- Security by Design and Privacy by Design
- Regulatory frameworks: The EU Digital Services Act, UK Online Safety Act, and Australia's eSafety Commissioner guidance
- Academic research
- Trust and safety expertise
- And many others

LIMITATION AND SCOPE

While we have attempted to include the most diverse sources and related works, ideas and research possible, this taxonomy will always be incomplete. It is a repository of current knowledge that requires ongoing expansion as new threats emerge and new approaches to mitigate these harms rise. **This taxonomy is not a techno-solutionist tool** that can solve systemic problems of

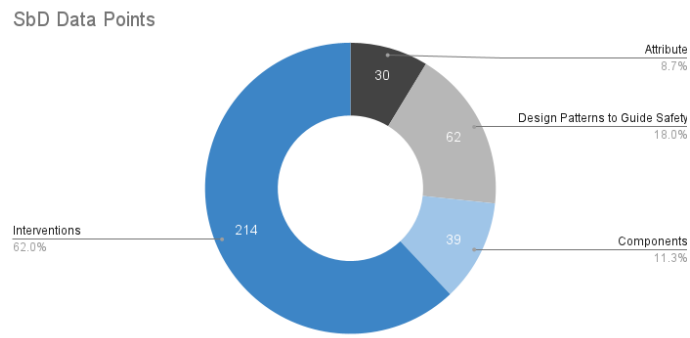
29 https://tfgbv.humane-intelligence.org/ NDI Landscape Tracker: TFGBV Interventions * tracker https://docs.google.com/spreadsheets/d/1SFJ-ZWryxYN9REXGCdM7dBjoBQM9-JHLJgC6osK0Kzo/

digital violence through design alone; **it is a reference source and a library.** Please note, there might be some systemic and critical interventions not included in the current taxonomy scope but oversight does not make those interventions any less necessary.

The taxonomy is organized into four distinct categories of attributes, design patterns, interventions, and components. Each category represents a specific type of knowledge or intervention from the technology facilitated violence space. The taxonomy and its visualizations are not hierarchical layers as each category is equally important; rather, each of the four broad categories represent different levels of information, types and implementation, from foundational values to concrete interventions.

1. Attributes (30 data points[30])

Attributes are the **core principles** that guide the intent and focus of Safety by Design design decisions. They represent the ethical and political foundation of the taxonomy. For example, often research and literature would reference a need that must be addressed by platforms. These needs, like 'agency', would have different ways to be implemented and designed depending upon the context of a problem, and the particular platform. But, these 'attributes' are principles that appeared again and again in the literature, and were needs that can guide design, and must be engaged with by platforms.



Data points that construct the Safety by Design Taxonomy.

Attributes define what "safety" means in a specific context and help evaluate whether a pattern or component aligns with that vision. For example, a design might be secure but not usable, or it might protect privacy while undermining user agency. The Attributes layer helps reveal those tensions.

Key Attributes include:

- **Agency:** The capacity of users to make meaningful choices about their safety, privacy, and participation
- **Usable:** Making safety features accessible and comprehensible to all users
- **Transparency:** Providing clear information about how systems work and what data is collected
- **Care:** Designing with empathy for users' emotional and psychological wellbeing

30 Data points references to the individual pieces of information that make up each category. Meaning, for the "attributes" category, we analysed and labeled over 30 different types of attributes in the taxonomy.

- **Consent:** Ensuring users have genuine control over how their data and content are used
- **Private:** Protecting users' personal information and digital traces
- **Intersectional:** Recognizing how multiple marginalized identities compound vulnerability
- **Accountability:** Creating mechanisms for when harm occurs
- And others

The Attributes layer reveals the values embedded in design decisions. When platforms claim to prioritize safety while implementing features that undermine user agency, or claim to enhance security while destroying privacy, the Attributes layer exposes these contradictions.

2. Design patterns to guide Safety (*62 data points*)

Design Patterns are strategies, behaviors, and/or interaction models used to guide the design of patterns for safety. Within design literature, design patterns are defined as "reusable/ recurring components which designers use to solve common problems in user interface design[31]." These design patterns are the blueprints, frameworks or approaches for tackling safety challenges. They sit between high-level values and technical implementation. Design patterns can offer a repeatable logic for how to address specific harms across contexts, but they are not as granular as "change this exact dashboard in this way."

Key Design Patterns include:

- Real-time-Intervention systems
- Community driven moderation
- Report tooling
- Technical safety measures
- Escalating penalty systems
- Rapid response
- Data minimization
- Compartmentalization
- And others

Design Patterns provide the conceptual bridge between abstract values and concrete features. They answer questions like: How do we operationalize consent in a messaging app? How do we balance transparency with usability? How do we create accountability mechanisms that don't require invasive surveillance?

3. Components (*39 data points*)

Components are concrete and specific tools, interface elements, or system architectures built into digital platforms to realize safety strategies. This is an actionable layer, where values and patterns are made concrete and usable. Components are the features and systems that users interact with: buttons, settings, algorithms and infrastructures that operationalize safety. Design

31 https://www.interaction-design.org/literature/topics/ui-design-patterns

components can also be thought of as 'user interfaces' or specific types of user experience design. Design components can be assessed by how well they embody the Attributes, address specific areas of harm, and implement Design Patterns.

Key Design Components include:

- Archiving your posts
- Quarantine content
- 2FA or 2 factor identification (security)
- E2EE or end to end encryption (confidentiality; security)
- Not collecting metadata (privacy)
- Warnings and flagged content
- Restricting users
- Changing profile names to create nicknames, new names, etc
- Creating a 'timeout for users'
- And others

The distinction between Design Patterns and Components is crucial but sometimes subtle. Components are specific instances (actual features, user interfaces, or systems) while Design Patterns are the conceptual approach that can be realized through various components.

When something might be both a Design Pattern to Guide Safety and a Component: There are cases when Design Patterns and Components operate simultaneously, depending on the implementation phase and perspective. This occurs during development when the conceptual strategy is very coupled with the concrete form that this might take, even affecting how some components may co-evolve or even reinforce each other through the design process. In such cases, the design pattern offers the guiding rationale and the component provides the instantiated expression. **The two layers are not mutually exclusive but interdependent.**

Ex. Automatic or AI moderation

- As a **Design Pattern:** Algorithmic or automated moderation frameworks to pre-empt harmful content.
- As a **Component:** A trained machine learning model that flags or removes messages in real time.

Ex. Panic Button

- As a **Design Pattern:** The idea of creating a "rapid exit" or creating a fast way to quickly change a lot of settings at once. This is applicable in many domains (ex. social platforms)
- As a **Component:** A specific clickable button that deletes history, triggers help, or triggers a wide variety of set features to do a particular action.

4. Interventions (*200+ data points*)

Interventions are the specific instances where patterns and components are applied. They embody how safety work is enacted in practice, showcasing specific

platforms, uses and particular contexts addressing defined harms.

Interventions are documented examples of Safety by Design in action. Starting from the NDI OGBV Tracker, these Safety by Design interventions include successful implementations worth replicating and failures worth learning from. The Interventions layer grounds the taxonomy in reality, showing real examples where abstract principles and design strategies actually play out when implemented. This layer is the most granular and context-specific. It includes platform-specific features, regulatory requirements, advocacy campaigns, grassroots tools and others.

A single Intervention can embody multiple **Attributes**, implement several **Design Patterns,** and utilize various **Components.** Conversely, a single Attribute (e.g. transparency) might be implemented and realized through dozens of different Design Patterns (e.g. nuanced privacy settings) and Components (e.g. the specific privacy settings created). This relational structure of an Intervention embodying Attributes, Design Patterns and Components, reflects the reality that Safety by Design is systemic, interdisciplinary and nuanced. For Safety by Design to really be implemented, all of these four categories must be engaged with and actually embodied within a platform's design.

Putting the Taxonomy into Practice:

For example, take this scenario: a content warning on sensitive materials that a user shares on a platform. This example illustrates how a Safety by Design approach spans all four categories, and how understanding these categories can contribute to evaluating a feature's Safety by Design quality.

- **Attributes:** Consent (users choose whether to view), Care (protecting psychological wellbeing), Agency (empowering user choice and protection)
- **Design Patterns:** Default protection (hiding potentially harmful content by default), real-time alerts (the act of adding a warning may prompt posters to reconsider sharing)
- **Components:** Warnings and flagged content (to identify sensitive material)
- **Interventions:** Introduce nudges or friction to discourage users from posting abusive content. (For example, Twitter introduced a new feature nudging users to read articles before sharing them. Using the expanded classifiers described above, platforms should consider similar measures to discourage the posting of abusive and harassing language in the first instance.)

Building the Taxonomy: Interactive Data Visualization

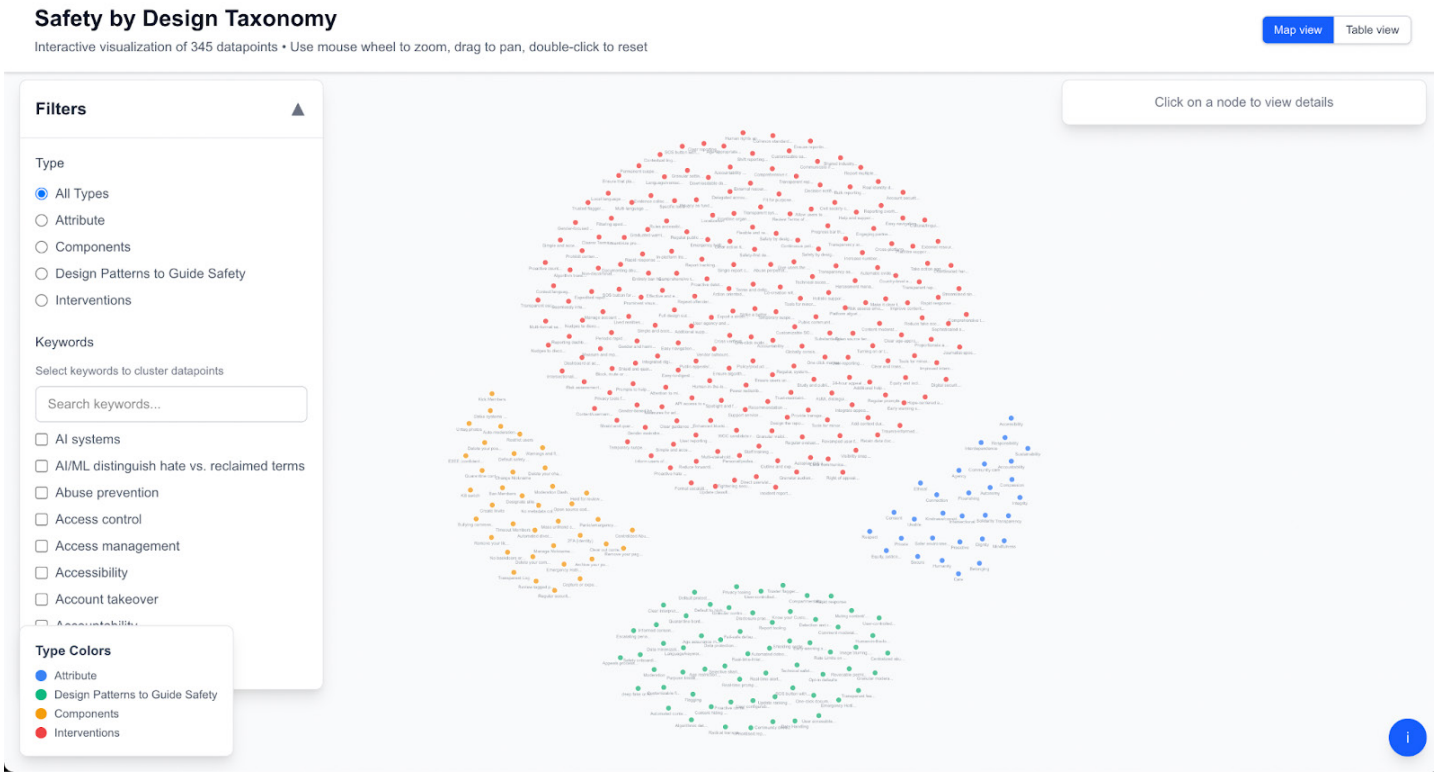
The Safety by Design Taxonomy has been developed as an interactive web-based data visualization accessible at: (https://sbd-taxonomy.vercel.app/). The visualisation transforms a series of interconnected data points organised in Google Spreadsheets (Table view) into an explorable network that allows users to dive into the taxonomy. The visualisation allows users to explore relationships between different elements, filter by specific attributes, design patterns, components or interventions and keywords.

The interactive format is essential because of the taxonomy's complexity - with 30 Attributes, 62 Design Patterns, 39 Components, and 200+ Interventions - being better represented than a static display. The visualisation allows any user to navigate the taxonomy according to their specific needs (e.g. platform designers evaluating their safety measures; policy makers interested in proposed or utilized solutions; advocates or researchers wanting to understand the scope and scale of Safety by Design interventions, etc).

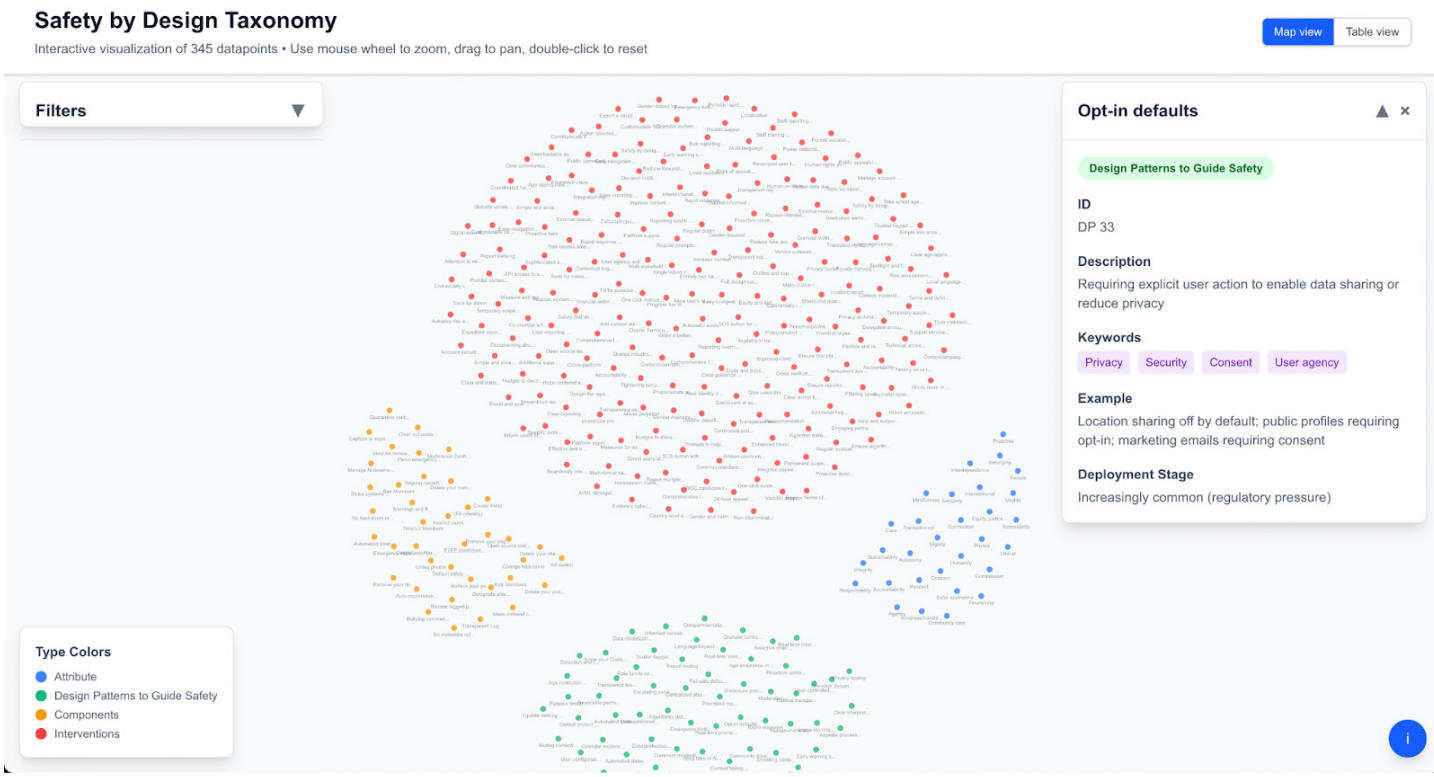
Lastly, **this tool is designed to evolve alongside the safety landscape as a living document.** As new threats emerge, interventions are developed, and platforms evolve, this taxonomy needs to be updated to reflect the current realities digital harm faces, transforming this to a living framework for accountability and improvement.

Key Features:

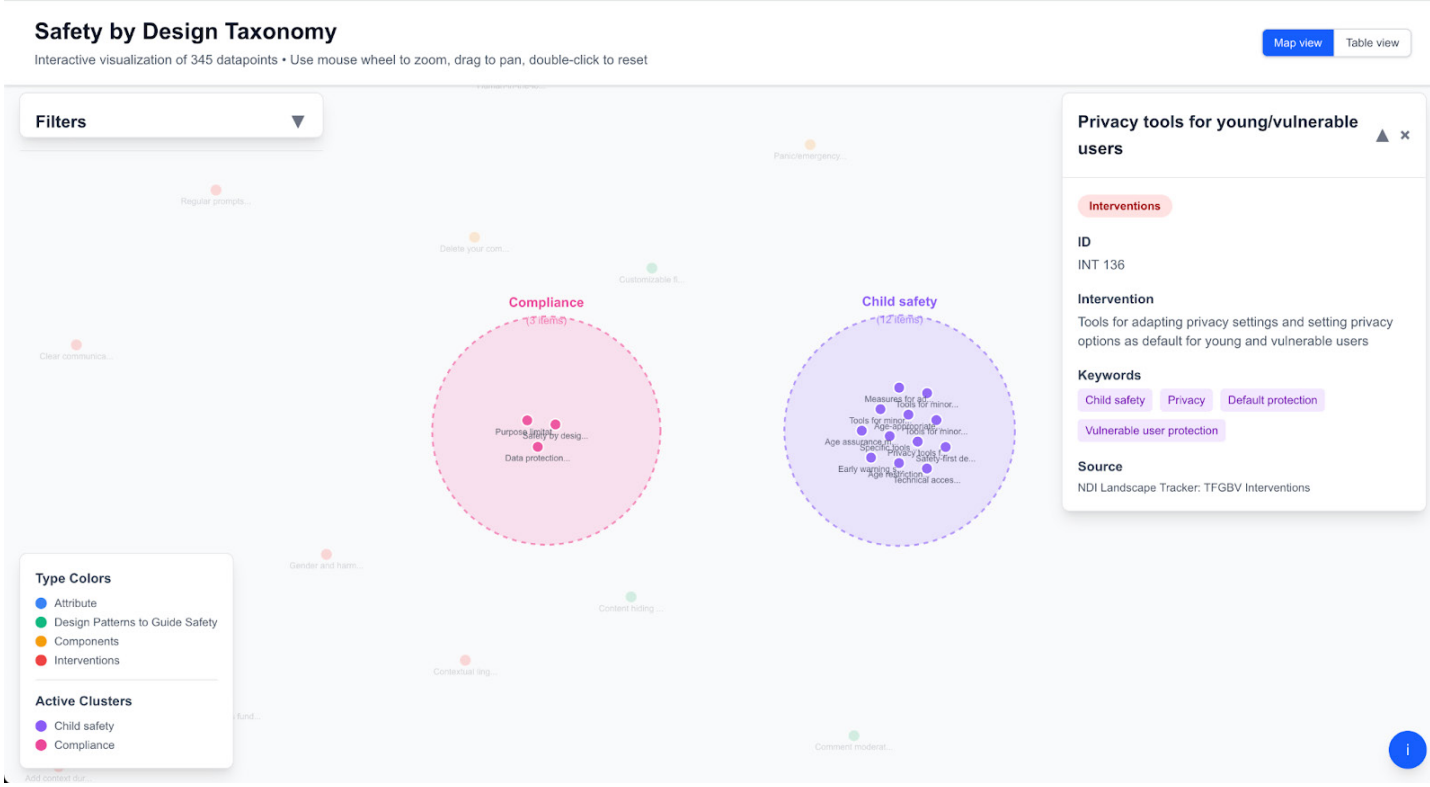
1. **Full network view:** Displays all the data points across the four taxonomy layers (Attributes, Design Patterns, Components, and Interventions) showcasing the interconnectedness of the data points within the system including related terms, links, and technology.
2. **Filtering capabilities:** Users can filter between all the layers that compose the taxonomy, keywords or deployment status.
3. **Metadata:** Each data point includes detailed information (Unique Identifier, Type Classification, Description, Deployment Stage, Keywords, Example and Source).
4. **Color-Coded Identifiers:** Data points are visually distinguishable by Layer in the taxonomy. When clustered using keywords each data point locates itself inside of the selected keyword concept and coloured accordingly.
5. **Future works:** Plans for making it searchable and exportable so users can search for specific terms, explore clusters of related interventions and cross reference the taxonomy.



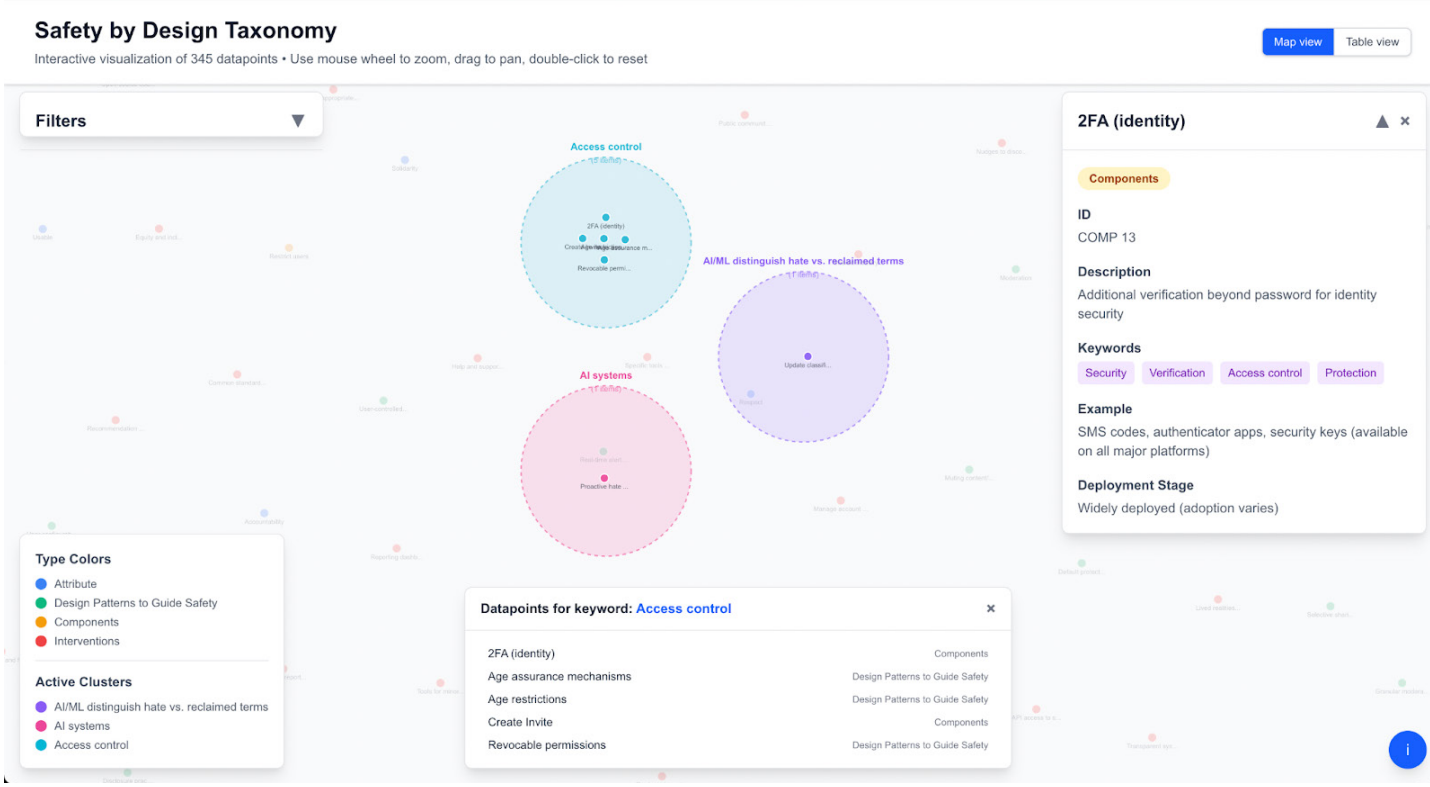
Main landing page visualisation - showcasing all the data points in the taxonomy organised by Layer/Type color.



Panel on the right - shows metadata of a selected datapoint.



Keyword clusters



Keyword clusters - activate information panel for each cluster on the bottom, when clicked the detail panel (right) activates with further information.

5. Case Studies of Online Violence and Leveraging Safety by Design

A note: this analysis on platforms' current trust and safety features was conducted between September and December 2025. Platforms' settings, features, toolings and policies are constantly shifting and updating, and, as such, the specific platform's safety toolings might have changed since the publication of this paper.

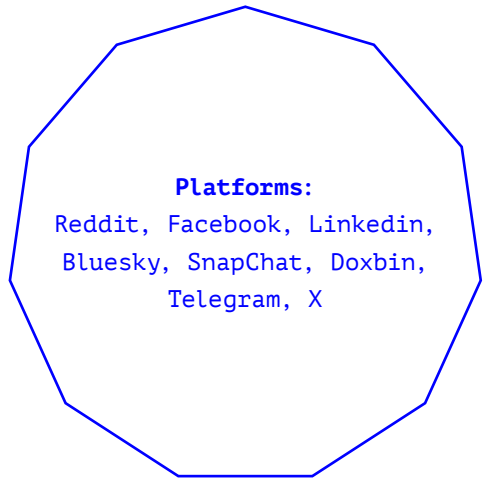
Implementation, interpretation, and analysis matters when engaging in a Safety by Design process. The following section presents 6 case studies developed to illustrate the application and implementation of Safety by Design, through an audit of the mentioned platforms, products and software in each case study. In each case study, our researchers accessed each platform, product, and/or software from the US, using VPNs, and creating dummy accounts or using personal accounts. Each feature mentioned in our case studies were also accessed via the platform or product and was then further cross referenced with documentation listed online from the company's websites, blog posts, documentation or manuals in terms of how that product should function, work, and where it could be accessed to catch any discrepancies between documentation and the feature accessed within the product, platform or software.

The six case studies represent anonymized scenarios built from real examples of harm, and are centered around specific platforms. In each case study, we outline what platforms currently offer in terms of harassment or harm mitigation and prevention, harassment support, security and privacy tooling and related policies. Each section then also suggests specific Safety by Design interventions, pulled directly from the Safety by Design taxonomy. Each suggestion for each case study can, in our expert opinions, be applied to other social media platforms. Each intervention, design component and pattern is transferable to other platforms, for as long as users can interact with each other, and socialize, then the harms highlighted in these case studies become transferable from platform to platform and then so too do the mitigation and prevention safety suggestions to address the specific harms.

Case Study 1: Public Figure Facing Network Harassment & Doxxing

The victim: Mayor Jennifer Walsh, a local politician in support of refugee and housing rights

Harm: Coordinated harassment attacks across platforms, coordinated digital harassment that encourages offline harassment at her events, doxing attempts, gendered harassment (including comments on her appearance), review bombing and targeted disinformation



Week 1: After announcing support for refugee housing, Mayor Walsh's official Facebook page receives appearance-based comments rather than policy discussion and feedback. Her official Linkedin page receives multiple LinkedIn messages questioning her "qualifications" with gendered undertones. Her official X (formerly Twitter) account replies include veiled threats like "we know where you work." Bluesky users "ratio" or overload her posts with coordinated messaging as comments[32]. Reddit's r/politics creates threads mocking her appearance. She reports each incident, but platforms respond that content "doesn't violate guidelines" given her stance as a public, elected figure.

Week 2-3, with harassment continuing to escalate: Her official Mayoral Facebook page gets review-bombed with false corruption allegations. Snapchat's Snap Map reveals her location at public events, attracting protesters. YouTube comment sections under city council streams fill with targeted sexual harassment direct at Mayor Walsh. TikTok records and clips her speeches that are then used as background sounds in misogynistic videos. Mayor Walsh uses tools like Facebook's Page Quality features, and limits Instagram comments to followers, mutes comments and users when possible on Meta products, but disinformation and harassment continues to spread faster than fact-checks after Meta replaced fact-checking with Community Notes in January 2025[33].

Week 4, and onward, harassment continues to propagate and become more severe, including doxxing attempts: Mayor Walsh's home address appears on Doxbin with her children's school information. WhatsApp groups share her personal phone number leading to 500+ calls daily. Telegram channels post AI generated and manipulated

32 https://www.bsky.social/about/blog/01-17-2025-moderation-2024

33 https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/

photos to create misinformation suggesting untruths like affairs. Local Facebook groups coordinate harassment at her home. Her teenage daughter's Snapchat receives threats. Mayor Walsh decides she needs 24/7 security, and implements a family social media lockdown. She decides to work with the FBI Election Crimes unit as a potential way to mitigate these harms; she also changes children's schools for their protection, and installs more home security. Because she is an elected official and considered a public figure, platforms provide less than adequate support for the frequency of harm she's facing. Mayor Walsh's harassment is similar to 41% of harassment incidents that span multiple platforms[34].

Safety by Design Suggestions to Mitigate Harms: In Mayor Walsh's case, she's in a particular policy predicament in which elected officials often have less harassment protections[35] in order to allow for public critique and disagreement of their policies. Platforms could create nuanced policy to support verified public figures and more nuanced protection across all platforms when 'public speech' crosses over into harassment or violence[36]. Platforms could consider creating community or 'family' account protection linking; implementing priority 24-hour response for elected officials which is something Meta has implemented; and emergency lockdown mode for all linked accounts during active threats.

34 https://www.pewresearch.org/internet/2021/01/13/the-state-of-online-harassment/

35 https://www.facebook.com/government-nonprofits/blog/safety-tools-and-policies-for-elected-officials-and-candidates

36 https://about.fb.com/wp-content/uploads/2021/10/Facebook_PolicyForum_-Attacks-on-Public-Figures.pdf

Intervention	Label	Example	Deployment Stage
Offer tools that allow users to report content that violates the companies' policies against abuse or mis- and disinformation and to control who can interact with their accounts.	User reporting tools and interaction controls	Universally available; quality and responsiveness vary	**Widely deployed (effectiveness varies)**

Platforms should revamp reporting features to be more user-friendly, responsive, and trauma-informed. Specifically, platforms should: Ensure clarity and consistency between reporting features and policies within platforms. When reporting, users usually select from preset choices to indicate how the content violates the rules. The language used in these preset choices to describe prohibited abusive tactics must be harmonized with the language used in platform policies. When using the reporting features, users should be able to quickly and easily access relevant policies so they can check, in real time, whether the content they are reporting likely violates those policies.	Revamped user-friendly reporting	Ongoing improvements; PEN America recommendations not fully implemented	**Continuous improvement but gaps still remain around user-friendliness, responsiveness and how they respond to trauma-informed frameworks.**
Create a formal, publicly known appeals or escalation process. When users report abusive content that they perceive to be in violation of platform policies and that content is not taken down, they should have accessible avenues to appeal or escalate their case. The appeal or	Public appeals process for report dismissals	Meta Oversight Board; YouTube appeals; general platform appeals (limited transparency)	**Implemented but there is limited transparency for the users. The appeal or escalation process should be formal, public, integrated into the reporting process, and available to all users who report harassment. Moreover, they should be able to amend their**

escalation process should be formal, public, integrated into the reporting process, and available to all users who report harassment. When users file an appeal or escalation, they should be able to amend their case, adding context or additional related abusive content.			**case, add more context or related abusive content.**
Platforms should create a transparent system of escalating penalties for all users, including warnings, strikes, temporary functionality restrictions, and suspensions, as well as content takedowns and account bans.This accountability system should be fully integrated into the primary user experience and clearly visible alongside policies governing user behavior.	Transparent escalating penalty system	YouTube strikes system; Twitter/X strikes; varying transparency	**Implemented but each platform has varying transparency and sometimes is not at all integrated in the user's primary user experience.**
Ensure that content moderation systems, including human moderators and algorithmic systems, are attuned to the needs of and the threats faced by women of color political candidates, in particular.	Content moderation attuned to intersectional threats	Proposed in research; limited public information on implementation	**Proposed/ unclear implementation**

Define effective policies for detecting and penalising repeat offenders, to stop the same abusers assuming new online identities after action taken such as suspension or de-platforming.	Repeat offender detection and penalties	Proposed by PEN America; YouTube channel manager access; Gmail delegate system (not harassment-focused)	**Proposed (limited creator tools exist)**
Create a formal, adequately resourced escalation channel for expediting appeals to address cases of malicious or inaccurate content takedowns and time-sensitive cases where a delay in restoring content or accounts could be harmful, for example during times of urgent political debate or crises. At the very least, this channel should enable an institution, such as a news outlet or civil society organization, to advocate with the platform on behalf of an individual, such as a journalist or human rights defender.	Formal escalation channel for time-sensitive cases	Twitter's former Trusted Flagger program; some journalist hotlines	**Limited/ declining (formerly Twitter)**
Regularly evaluate and publicly report on how social media platforms fuel discrimination, bias, and hate, and then make product or policy	Regular evaluation and public reporting on discrimination	DSA requires risk assessments; limited public comprehensive evaluations	**Regulatory requirement (EU), limited voluntary action**

improvements based on these evaluations. Despite what tech companies claim, having millions or even billions of users is not among the primary problems—defective policies, bad products, and subpar enforcement are, along with a business model that rewards optimizing for engagement.			

MORE DESIGN PATTERNS FROM THE TAXONOMY TO GUIDE SAFETY:

Design Pattern	Example	Deployment Stage
Real-time-Intervention systems, including: a cross platform doxxing alert system	Instagram's "Are you sure you want to post this?" nudge for offensive comments; Twitter/X reply warning system	**Product integration (major platforms) with no cross doxxing alert systems**
Automated detection systems: which could be proactive scanning and removal of residential addresses which would help victims like Mayor Walsh and her family	YouTube Content ID; Meta's automated hate speech detection; Twitch AutoMod	**Deployed but with limited features - there are automated detection systems but not ones able to flag addresses**
User-controlled visibility	Twitter protected tweets; Instagram private accounts; Facebook friend lists and selective sharing	**Widely deployed**

Default protection settings	Instagram hiding sensitive content by default; Snapchat friend lists hidden for minors	**Implemented (varying by platform)**
Better Data Protection Integration	GDPR-compliant cookie banners; right to deletion features	**Regulatory requirement (deployed)**
SOS button with live support ala the “emergency lockdown mode” listed above or a panic button	“emergency lockdown mode” listed above or a panic button	**Research/prototype phase**
Rapid response	Twitter’s former trusted flagger expedited review	**Limited implementation (declining)**
Report tooling	Instagram’s multi-step reporting flow; YouTube’s video/channel reporting; X bulk tweet reporting	**Widely deployed (quality and tooling granularity varies)**
Privacy tooling	Facebook privacy checkup; Instagram close friends list; Twitter protected tweets	**Widely deployed**
Escalating penalty systems	Twitter/X strike system; YouTube Community Guidelines strikes; Twitch suspension escalation	**Widely deployed (strike system is widely deployed but its difficult for users to know how it’s deployed or even know their strikes)**

Centralized abuse dashboard	YouTube Copyright Match Tool dashboard (exists for IP, not harassment); PEN America recommendation	**Proposed (not implemented for harassment)**
Technical safety measures	Two-factor authentication(2FA) (widely available); end-to-end encryption (E2EE) (Signal, WhatsApp); login alerts	**Some widely deployed (2FA) but others like E2EE not at all (Signal exemplar)**
Trusted flagger programs	YouTube Trusted Flagger Program; Twitter's former human rights defender program; Global Internet Forum to Counter Terrorism (GIFCT)	**Deployed but still limited**
Flagging	User reports; automated flagging; trusted flagger priority queues	**Widely deployed**
Shielding content	Block Party lockout folders (discontinued); Reddit Harassment Filter for moderators	**Limited deployment (third-party tools, some platform features)**
Data minimization	Privacy by design principles; limited data retention	**Regulatory requirement (enforcement varies)**
Selective sharing	Facebook audience selector; Instagram close friends stories; LinkedIn connection-only posts	**Widely deployed**

Appeals processes	YouTube copyright appeals; Meta's Oversight Board; general content moderation appeals	**Widely deployed (effectiveness varies)**
Early warning system	A user-facing early warning system that detects coordinated harassment campaigns is currently missing.	**Limited implementation to detect harm before they exist.**

MORE SPECIFIC DESIGN COMPONENTS FROM THE TAXONOMY:

Design Components	**Deployment Stage**
Review tagged photos, which X, Instagram and Facebook currently offer; Bluesky allows for the tagging of users in general posts but does not allow photos to be tagged yet.	**Widely deployed**
Delete comments and reactions, as platforms like YouTube, Instagram and Facebook allow for page owners, like Mayor Walsh and her team, to delete comments or those that posted the content to delete comments. But, generally, across all platforms, a user cannot delete comments on posts or content that are not theirs.	**Mostly manual**
Manage Nicknames	**Widely deployed (community platforms)**
Bullying comment filters (which differs from a quality filter)	**Implemented (major platforms)**

Clear out or delete content, such as comments, in bulk	**Regulatory requirement (GDPR), limited bulk tools**
No metadata collection	**Limited implementation (Signal exemplar)**
Regular security audits	**Common (large platforms), quality varies**
Warnings and flagged content	**Widely deployed**
Default safety settings automatically enabled	**Partial implementation (varies by platform and user age)**
Capture or export evidence of abuse (screenshots, logs) easily	**Limited implementation (manual screenshots standard)**
“Held for review” queues moderation	**Selectively deployed**
Designate allies who can help monitor, block, report, document abuse on their behalf	**Limited (YouTube creators), proposed comprehensively**
Restrict users	**Implemented (Instagram)**
Kick Members	**Widely deployed (community platforms)**

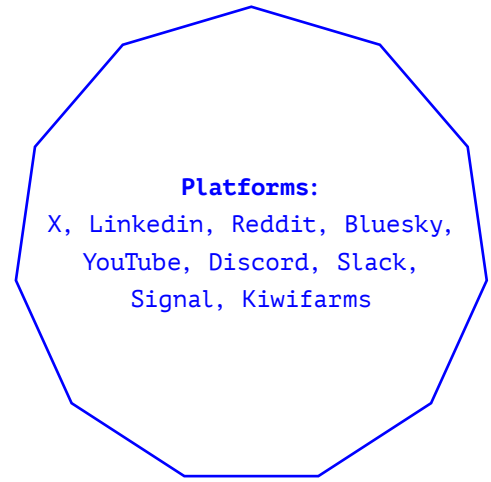
Ban Members	**Widely deployed**
Timeout Members	**Widely deployed (community platforms)**
Strike systems to raise accountability for repeat offenders	**Widely deployed (transparency varies)**
Centralized Abuse Dashboard	**Proposed (not implemented for harassment)**
Quarantine content	**Limited implementation**
Moderation Dashboard	**Widely deployed (moderator/admin tools)**
Panic/emergency feature that escalates support and connects users with real time human assistance	**Research/prototype phase (DV apps only)**

Case Study 2: Semi-Public Figure - Marginalized Identity Harassment

The victim: River Chen, Climate Journalist & Trans Rights Advocate

Harm: deadnaming (using a trans or nonbinary person's 'former' but not chosen and preferred name), misgendering, gendered harassment related to their trans identity, coordinated harassment campaigns across multiple platforms, doxxing attempts targeted at their personal life, and digital stalking

Week 1: After publishing an article on environmental racism, River faces intersectional and gendered harassment. X users deliberately misgender them despite their pronouns listed in their bio. LinkedIn profiles send messages to River, describing their trans identity as "mental illness." Reddit users screenshot River's transition timeline and post these screenshots with mocking and harmful comments. Different Bluesky users create lists that label River as a "woke propagandist." They report but platforms say misgendering "isn't harassment[37]"; X/Twitter only enforces "where required by local laws". Meta currently has no deadnaming protection[38]. Bluesky does not allow users to remove themselves from a list that someone else has created and added them to[39].



Month 1, coordinated campaigns continue: Coordinated campaigns across multiple platforms continue to use River's deadname that was found through court records. River's transition before and after photos spread on Instagram with transphobic captions, and without River's consent. YouTube creators create "exposed" videos about River's "real identity, creating disinformation about River in the process. Multiple discord servers coordinate mass-reporting campaigns about River's work and social media profiles as a way to get them kicked off social media. River's work Slack receives infiltration attempts. Different Signal groups share River's therapist's information as a form of harassment and intimidation. The website Kiwi Farms creates a 200-page thread documenting River's entire life. What River is facing isn't new for LGBTGIA+ activists and journalists; according to GLAAD, 918 anti-LGBTQ incidents occurred in 2024, averaging 2.5 daily[40].

37 https://www.techpolicy.press/x-s-updated-misgendering-and-deadnaming-policy-should-concern-all-social-media-users-and-believers-in-democracy/
38 https://glaad.org/smsi/report-meta-fails-to-moderate-extreme-anti-trans-hate-across-facebook-instagram-and-threads/
39 https://www.reddit.com/r/BlueskySocial/comments/1k38uev/removal_from_moderation_lists/
40 https://glaad.org/releases/glaads-alert-desk-releases-new-data-on-rising-anti-lgbtq-hate-and-extremism/

Month 2 and onward, harassment continues to escalate: Deepfakes of River using their pre-transition photos appear on porn sites. Their employer faces boycott campaigns on LinkedIn. River's parents are targeted and receive Facebook messages about their "failed parenting" in regards to River's gender identity and journalism. River's pharmacy gets false DEA reports about River's hormone prescriptions. River's dating app profiles are screenshotted and shared across Reddit and X. River decides to legally change names in all jurisdictions, use virtual mailbox services, and implement compartmentalized digital identities via MySudo. They also start working with GLAAD's Social Media Safety Program[41] and consider relocating within their city to a new apartment. What River is facing is a systemic issue other nonconforming journalists face; for example, 90% of female and gender non-conforming journalists identify online harassment as the biggest threat to their work and safety[42].

Safety by Design Suggestions to Mitigate Harms: Platforms could create better policies for trans and nonbinary users, including more nuanced policy on deadnaming and misgendering as bullying or harassment. Platforms could create better protected status for journalists covering marginalized communities as opposed to treating journalists as 'public figures'. Platforms can implement proactive removal of before/after photos without consent, and create cross-platform coordination for identity-based harassment. Platforms could create better support for LGBTQIA+ users such as direct pipeline for users to LGBTQ+ support organizations.

41 https://glaad.org/smsi/lgbtq-social-media-safety-program/
42 https://cpj.org/2019/09/canada-usa-female-journalist-safety-online-harassment-survey/

Intervention	Label	Example	Deployment Stage
Platforms should support the creation of promising new third-party tools designed to counter online abuse—especially those built by and for women, BIPOC, and/or LGBTQIA+	Platform support for third-party anti-abuse tools	Formerly available; X/ Twitter API pricing decimated ecosystem; Facebook/Meta limited access	**Declining (API restrictions increasing)**

technologists with firsthand experience of online abuse—by investing in R&D and providing access to application programming interfaces (APIs), data, and other relevant information.			
Conduct gender-focused human rights impact assessments on platforms.	Gender-focused human rights impact assessments	Proposed by civil society; no industry-wide implementation	**Proposed (not implemented)**
Contribute to and use a shared industry global lexicon repository on gender issues.	Shared industry gender lexicon	Proposed by PEN America; TRFilter (discontinued); Instagram report history (limited)	**Proposed (manual screenshots standard)**
Social media companies should undertake a human rights and gender discrimination impact assessment that identifies, prevents and mitigates any negative impact of their operations on the rights to freedom of expression, privacy, participation and non-discrimination of women and women journalists.	Human rights and gender impact assessments	UN Guiding Principles requirement; DSA risk assessments; limited public examples	**Regulatory requirement (limited transparency)**
Clearly articulate policies that prohibit content that harasses or abuses someone on the basis of gender or race.	Clear gender/ race harassment policies	Most platforms have policies; enforcement inconsistency noted	**Widely deployed (enforcement inadequate)**

Regulated services which outsource any part of their business, including moderation of content, applications, GIFs, images, or any other content or tools, including safety tech, should ensure the vendor adheres to the social media provider's Terms of Service and Community Standards, and where necessary take action to enforce those standards, and that they have employee and mental health protection policies in place that adhere at least to the same standard.	Vendor outsourcing standards	Some platform commitments; investigative journalism reveals gaps	**Stated policies (enforcement questioned)**
Measure the prevalence of gendered abuse and share data through corporate transparency reports.	Measure and report gendered abuse prevalence	Limited gender-disaggregated data in transparency reports; proposed by civil society	**Limited implementation**

MORE DESIGN PATTERNS FROM THE TAXONOMY TO GUIDE SAFETY:

Design Pattern	Example	Deployment Stage
Centralized abuse dashboard	YouTube Copyright Match Tool dashboard (exists for IP, not harassment); PEN America recommendation	**Proposed (not implemented for harassment)**
Rapid response	Twitter's former trusted flagger expedited review; account lockdown during active threats	**Limited implementation (declining)**
User accessible moderation tools	Discord server moderation; Facebook group admin tools; subreddit moderator controls	**Widely deployed**
Informed consent	GDPR consent banners; app permission requests with explanations; Instagram's data usage notifications	**Regulatory requirement (quality varies)**
Data Handling	End-to-end encryption (Signal, WhatsApp); minimal data retention policies	Varying deployment
Report tooling	Instagram's multi-step reporting flow; YouTube's video/channel reporting; X bulk tweet reporting	**Widely deployed (quality varies)**

Shielding content	Block Party lockout folders (discontinued); Reddit Harassment Filter for moderators	**Limited deployment (third-party tools, some platform features)**
Trusted flagger programs	YouTube Trusted Flagger Program; Meta's partnership with counter-terrorism orgs; Twitter's former human rights defender program	**Implemented (varying commitment)**
Escalating penalty systems	Twitter/X strike system; YouTube Community Guidelines strikes; Twitch suspension escalation	**Widely deployed (transparency varies)**
Data minimization	Privacy by design principles; limited data retention; aggregate statistics instead of individual tracking	**Regulatory requirement (enforcement varies)**
User-controlled visibility	Twitter protected tweets; Instagram private accounts; Facebook friend lists and selective sharing	**Widely deployed**
Appeals processes	YouTube copyright appeals; Meta Oversight Board; general content moderation appeals	**Widely deployed (effectiveness varies)**
Flagging	User reports; automated flagging; trusted flagger priority queues	**Widely deployed**

MORE SPECIFIC DESIGN COMPONENTS FROM THE TAXONOMY:

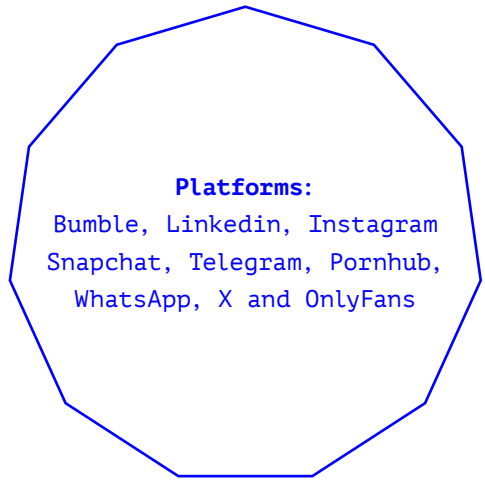
Design Components	Deployment Stage
Default safety settings automatically enabled	Partial implementation (varies by platform and user age)
Capture or export evidence of abuse (screenshots, logs) easily	Limited implementation (manual screenshots standard)
Auto-moderation filters	Widely deployed
Clear out content in bulk	Regulatory requirement (GDPR), limited bulk tools
Bullying comment filter	Implemented (major platforms)
Emergency Hotline	External partnerships exist; in-platform proposed
Panic/emergency feature that escalates support and connects users with real time human assistance	Research/prototype phase
Designate allies who can help monitor, block, report, document abuse on their behalf	Limited (YouTube creators), proposed comprehensively
Moderation Dashboard	Widely deployed (moderator/admin tools)

No metadata collection (private)	**Limited implementation (Signal exemplar)**
Regular security audits (accountability)	**Common (large platforms), quality varies**
Warnings and flagged content	**Widely deployed**
“Held for review” queues moderation	**Selectively deployed**
Restrict users	**Implemented**
Ban Members	**Widely deployed**
Strike systems to raise accountability for repeat offenders	**Widely deployed (transparency varies)**
Centralized Abuse Dashboard	**Proposed (not implemented for harassment)**
Quarantine content	**Limited implementation**

Case Study 3: Dating Violence & Image-Based Sexual Abuse

The victim: Sarah Martinez, a dating app user facing escalating abuse

Harm: harassment, stalking, and unsolicited visual depiction of intimate parts/image-based sexual abuse



Week 1: Sarah matches with Tom on Bumble. After one date, she politely declines a second date. Tom finds her Instagram through reverse image search, and begins sending her roses emojis on every post. Her LinkedIn receives connection requests from fake recruiter profiles that seem to be Tom. Snapchat shows him constantly viewing her location. Sarah blocks Tom everywhere, but new accounts and fake profiles keep appearing, and reengaging with her. Sarah reports Tom's Bumble account to Bumble for harassment and blocks him. Bumble eventually bans his account.

Week 2-3, harassment continues and spreads to other platforms: Sarah receives unsolicited visual depictions of intimate parts (e.g. dick pics) from fake profiles that flood her Instagram DMs. Instagram automatically blurs these images, and Sarah reports each image and blocks each fake account one by one. Her WhatsApp number receives explicit videos from international numbers; Sarah blocks each new number and reports the harmful content. She changes her WhatsApp settings to block unknown numbers[43]. Numerous Telegram accounts start to share her Bumble photos in "rate local women" groups. Reddit's local city subreddit receives posts with her photo asking "anyone know this girl?" Sarah starts to receive Signal messages from unknown numbers with intimate questions, and she changes her Signal settings to disable messages from non contacts in her phone[44]. She reports to Telegram and Reddit but gets automated responses about "not violating community guidelines." Some of her Instagram reports are viewed as harassment but others are not, leaving Sarah confused.

Month 1 and onward, with harassment escalation including the creation of deepfakes: Tom creates deepfake pornography of Sarah using her social media photos, and uploads the deepfakes to websites like Pornhub, XVideos, and OnlyFans under her real name. The videos then spread to Telegram channels with 50,000+ members. Sarah's employer receives anonymous emails with video

43 https://faq.whatsapp.com/3379690015658337/

44 https://support.signal.org/hc/en-us/articles/9932632052378-How-to-protect-yourself-on-Signal

links to the deepfake pornography. Sarah's immediate family members get WhatsApp messages with the same content. Google search results for her name now also show the deepfake videos. Sarah must file DMCA takedowns; often platforms take an average of 3 weeks to respond to non-consensual intimate image reports, with porn sites often never responding[45]. Sarah starts work with Cyber Civil Rights Initiative[46]. She also hires a reputation management firm. She begins to pursue criminal charges in jurisdictions that recognize image-based sexual abuse. Sarah decides to change jobs due to workplace harassment she receives.

45 https://www.thorn.org/blog/what-the-2024-ncmec-cybertipline-report-says-about-child-safety/
46 https://cybercivilrights.org/ccri-safety-center/

Safety by Design Suggestions to Mitigate Harms: Platforms and websites could create better tools to understand sexual harassment cases, and more easily allow for users to share their reports across platforms. Platforms and websites could implement mandatory 24-hour removal for non-consensual content and create a direct reporting pipeline for those impacted by image-based sexual violence. Platforms and websites could research the efficacy and safety implications of using technology innovations like image hashing or proactive deepfake detection using biometric markers by focusing on what is the most safe, secure and most respectful for the victim, their images and the victim's privacy. Platforms and websites, like dating websites, could also create verified consent systems for those that want to share intimate content with other, consenting users.

Intervention	Label	Example	Deployment Stage
Providers must have transparent processes that highlight the journey of a user report, and opportunities for regular updates and communication about action taken. If action is not taken to remove content, there must be a clear explanation as to why, together with signposting to relevant services who can help and	Transparent report journey and status updates	UK Online Safety Act requirement; Instagram report status; proposed comprehensive systems	**Regulatory requirement (UK), partial implementation**

support and potentially appeal a decision (such as the Revenge Porn Helpline or Report Harmful Content Service).			
Actors cannot take advantage of new or emerging tools to cause harm to women and girls. For instance: • deep fake or audio • visual manipulation materials. • nudification technology. • bots and bot networks. • content embedded from other platforms and synthetic features such as gifs, emojis, hashtags. • other new technology.	Prevention of emerging harm technologies	X/Twitter deepfake labeling; Meta AI-generated content labels; some platforms ban nudification apps; bot detection systems	Emerging (labeling more common than removal)
Regulated services should risk assess the tools for the creation of content - this includes but is not limited to bots (including chatbots), bot networks, deepfake or audiovisual manipulation materials, content embedded from other platforms and synthetic features such as gifs, emojis, hashtags.	Risk assessment of content creation tools	UK Online Safety Act requirement; limited public assessments	**Regulatory requirement (UK), implementation unclear**

Build shields that enable users to proactively filter abusive content (across feeds, threads, comments, replies, direct messages, etc.) and quarantine it in a dashboard, where they can review and address it with the help of trusted allies.	Shield and quarantine dashboard	Proposed by PEN America; Block Party (discontinued); Reddit Harassment Filter (moderators only)	**Proposed (limited moderator tools exist)**
Initiate platform-platform cooperation, since online violence often jumps across platforms and exploits the weaknesses of each.	Cross-platform cooperation against abuse	Limited examples; GIFCT for terrorism; proposed for harassment	**Limited (terrorism-focused, not harassment)**
Offer all three mechanisms—blocking, muting, and restricting—and apply each of these mechanisms consistently across all different forms of communication, including comments, DMs, tags, and mentions, and across desktop and mobile apps, regardless of device or browser type.	Consistent blocking/muting/restricting across platforms	Instagram restrict feature most comprehensive; inconsistent across platforms and features	**Fragmented (Instagram leading)**
Platforms should use the full suite of design elements (nudges, labels, contextual clues, etc.) to communicate clearly and consistently with users across all available channels (within platform, via email, etc.) about	Clear communication about violations and penalties	Automated notifications; varying clarity and helpfulness	**Widely deployed (quality varies)**

what rule has been violated, current and potential future penalties, and next steps, including how to appeal.			
Regulated services must have in place processes to ensure that where machine learning and artificial intelligence tools are used, they operate in a non-discriminatory manner and that they are designed in such a way that their decisions are explainable and auditable. For instance, technology to remove sexualised pictures must not remove photos of breast feeding. A platform provider should consider the way in which AI and machine learning systems and/ or human moderators will distinguish between hateful and harmful content, reclaimed terms used by particular groups, and that of 'counter speech', minimising the risk of blocking or limiting legitimate use of terms within certain online communities and counter speech.	Non-discriminatory AI/ML with explainability	UK Online Safety Act requirement; DSA algorithmic transparency; limited public auditing	**Regulatory requirement (UK/EU), implementation unclear**

MORE DESIGN PATTERNS FROM THE TAXONOMY TO GUIDE SAFETY:

Design Pattern	Example	Deployment Stage
Image blurring, especially for image based sexual violence and non-solicited intimate images	Bumble's AI blur + reporting for nudes	Emerging
Automated content monitoring	Facebook's proactive hate speech detection; CSAM scanning systems	**Widely deployed (varies by harm type)**
Informed consent	GDPR consent banners; app permission requests with explanations; Instagram's data usage notifications	**Regulatory requirement (quality varies)**
Content hiding mechanisms	Twitter content warnings; Instagram sensitive content blur; Reddit NSFW tags	**Widely deployed**
Deepfake or AI generated content flagging	X's Community Notes for AI content; Meta's AI labeling; detection research tools	Emerging (labels more common than removal)
Algorithmic detection systems	PhotoDNA for CSAM; Amazon Rekognition	**Widely deployed**
Muting content/users	X (Twitter) mute accounts, keywords, hashtags; Instagram mute posts/stories; Facebook "snooze" for 30 days	**Widely deployed**

Widely deployed	GDPR-compliant cookie banners; right to deletion features; data portability tools	**Regulatory requirement (deployed)**
Proactive content filtering	Instagram automatic offensive comment hiding; Snapchat nudity detection; spam filters	**Widely deployed**
Detection and response	AirTag stalking alerts; suspicious login detection and account lockdown	**Implemented (evolving)**
Shielding content	Block Party lockout folders (discontinued); Reddit Harassment Filter for moderators	**Limited deployment (third-party tools, some platform features)**
Centralized abuse dashboard	YouTube Copyright Match Tool dashboard (exists for IP, not harassment); PEN America recommendation	**Proposed (not implemented for harassment)**
User accessible moderation tools	Discord server moderation; Facebook group admin tools; subreddit moderator controls	**Widely deployed**
Appeals processes	YouTube copyright appeals; Meta Oversight Board; general content moderation appeals	**Widely deployed (effectiveness varies)**
Flagging	User reports; automated flagging; trusted flagger priority queues	**Widely deployed**

Trusted flagger programs	YouTube Trusted Flagger Program; Twitter's former human rights defender program; Global Internet Forum to Counter Terrorism (GIFCT)	Implemented (varying commitment)
Report tooling	Instagram's multi-step reporting flow; YouTube's video/channel reporting; X bulk tweet reporting	**Widely deployed (quality varies)**
Revocable permissions	App permission revocation (iOS/ Android); GDPR right to deletion; OAuth token revocation	**Standard (platform dependent)**
Data minimization	Privacy by design principles; limited data retention; aggregate statistics instead of individual tracking	**Regulatory requirement (enforcement varies)**
Real-time alerts	Login from new device alerts; AirTag "traveling with you" alerts; unusual activity notifications	**Widely deployed**

MORE SPECIFIC DESIGN COMPONENTS FROM THE TAXONOMY:

Design Components	Deployment Stage
Strike systems to raise accountability for repeat offenders	**Widely deployed (transparency varies)**
Capture or export evidence of abuse (screenshots, logs) easily	**Limited implementation (manual screenshots standard)**
Restrict users	**Implemented**
Ban Members	**Widely deployed**
Timeout Members	**Widely deployed (community platforms)**
Warnings and flagged content	**Widely deployed**
Quarantine content	**Limited implementation**
“Held for review” queues moderation	**Selectively deployed**
Default safety settings automatically enabled	**Partial implementation (varies by platform and user age)**

Moderation Dashboard	**Widely deployed (moderator/admin tools)**
Centralized Abuse Dashboard	**Proposed (not implemented for harassment)**
Clear out content in bulk	**Regulatory requirement (GDPR), limited bulk tools**
No metadata collection (private)	**Limited implementation (Signal exemplar)**
Bullying comment filter	**Implemented (major platforms)**
Auto-moderation filters	**Widely deployed**
Delete your comments and reactions	**Mostly manual**

Case Study 4: Intimate Partner Digital Stalking

The victim: Maria Thompson, an intimate partner violence survivor

Harm: Stalking Maria via her apps that share location, targeting her child and family members on different apps like Roblox, digital stalking of her accounts via her friends and family, attempts to log in to her banking apps,and physical stalking by visiting friends and placing AirTags on her car.

Month 1, harassment starts: After leaving Jake, Maria starts to notice that he views all her Instagram stories within minutes using his account, and fake accounts. She starts to block him on all social media platforms, including accounts she believes are his. Her TikTok's algorithm keeps showing his videos despite her blocking him. Jake begins digital stalking Maria by following her public Venmo transactions that reveal her payments; Maria realizes these transactions defaulted to be shared publicly and changes her profile to make all past, present and future transactions private[47]. Maria notices they are still linked on Spotify and can see him listening to their songs as a public activity, she blocks him and reports him. Spotify does not find this to be harassment. She switches all of her accounts to private but discovers he screenshots everything through mutual friends' accounts. She tries to report this but platforms find this to not be harassment as she has no 'evidence' of the screenshots being shared.

Month 2, harassment continues to escalate: Jake posts intimate photos on specific sub-Reddits with enough details that Redditors identify her. While scrolling through Facebook Marketplace, Maria notices a new profile of Jake's with this profile selling his ex-girlfriend's belongings, which are her items he has kept. She reports to Facebook Marketplace but it is not found to be harassment as she lacks evidence of owning the items and does not have a police report about the items. Some of her friends report seeing Jake drive by their houses or stop at their front doors via their Ring doorbell camera, which concerns Maria about her safety and her friend's safety. Maria then discovers AirTags hidden on her car, most likely to track her movements. She discovers them via her own iPhone which lets her know an AirTag is close by. Had Maria owned an Android device, she might not

47 https://help.venmo.com/cs/articles/changing-payment-privacy-hiding-past-payments-vhel191

have discovered the AirTags. Her banking apps email her about login attempts, which Maria deduces is most likely Jake. She enables all privacy settings but it's hard to prevent and mitigate all of Jake's actions as platforms don't recognize stalking patterns across services.

Month 3 and onward, the harassment escalates even more: Maria discovers stalkerware on her phone, most likely from Jake. Her child's gaming accounts like Roblox and Minecraft receive messages from "dad's friend", leading Maria to grow more and more concerned. Maria and her son have a talk, and Maria starts to monitor his online behavior as a safety precaution. Maria obtains a restraining order and starts documenting all of the harassment and harm she's faced. Maria does not post anymore on social media, and her social media accounts, while heavily locked down, still receive numerous fake profile requests. She asks her friends and family to not post any photos or information about her or her son, but sometimes family members forget. Maria tries to request the photos be removed and deleted but since they are not 'her photos' and instead are 'photos of her', the platforms keep the photos up.

Safety by Design Suggestions to Mitigate Harms: Platforms could implement survivor support tools like support modes that hide all activity from specific accounts, and better implement nuanced policy to support stalking and intimate partner violence survivors, and educational materials for friends and family to better support intimate partner violence survivors. Platforms could consider creating better stalking pattern detection across services like Meta, Google, etc. Tools like emergency lockdown or panic buttons could be activated more easily across all platforms, with consent and implementation from the victim.

Intervention	Label	Example	Deployment Stage
Specific tools in place for users under 18. This could include: • Tools to stop children from receiving unsolicited messages from adults	Tools for minors: unsolicited adult messages prevention	TikTok DM restrictions for teens; Instagram message restrictions; adult notification systems	**Implemented (major platforms)**

• Notifications to make an adult messaging a child aware of the policies of the service in relation to communication with children • Notifications to ask a child if they know who is messaging them and to explain what children can do if they are confused or made to feel uncomfortable by it			
Platforms should create an Emergency hotline (phone and/or chat) that provides users facing extreme online abuse (such as cyberstalking or a coordinated mob attack) with personalized, trauma-informed support in real time.	Emergency hotline with trauma-informed support	External crisis line partnerships; proposed in-platform live support	**External partnerships; in-platform proposed**
Account security systems which enable survivors of abuse, who are hacked and locked out, to recover their accounts.	Account security systems for hacked survivors	Twitter/X “hacked account” recovery flow; Facebook account recovery for compromised accounts	**Implemented (quality/ speed varies)**
Enable users to automatically capture all relevant publicly available data for content that is flagged as abusive via user	Automatic evidence capture with legal standards	No platform has implemented this comprehensively to legal evidentiary	**Not implemented to our knowledge, not implemented to legal standards.**

muting, restricting, blocking, or reporting, as well as proactive content filtering. To ensure that the feature meets evidentiary requirements for legal proceedings, platforms need to work with civil society, law enforcement, and legal experts to create documentation standards in each country and jurisdiction where they operate.		standards.	
Expand tools and services for targets of harassment, ensuring that social media platforms are easily accessible and effective for those facing or fearing an online attack. Such tools can include the ability to batch report violative content and block multiple perpetrators simultaneously, an abuse-report tracking portal, and real-time rapid-response support.	Comprehensive tools for harassment targets	Partially implemented across platforms; no comprehensive suite	**Fragmented implementation**
Enable users to assemble rapid response teams and delegate account access, so that trusted allies can jump in to provide targeted assistance, from mobilizing supportive communities to helping	Rapid response teams with delegated access	Proposed; requires platform-civil society-legal expert collaboration	**Proposed (not implemented)**

document, block, mute, and report abuse. Enable users to proactively designate a limited number of trusted allies to serve as a rapid response team—a small network of trusted allies who can be called upon to work together or individually to monitor, report, and document abuse that is publicly visible and rally a broader online community to help.			
Offer users the ability to grant specific members of their rapid response team access to their accounts, akin to the "delegate" system available on Gmail. These delegates could assist with tasks that require direct account access, such as blocking, muting, and reporting abuse in DMs. Users should be able to control the level of access their delegates have (to public feeds versus private DMs, for example). Twitter should integrate its "teams" feature, which is currently available only through TweetDeck, more directly into the primary user experience and empower users to specify exactly which	Delegated account access ("rapid response teams")	Proposed comprehensively; YouTube teams feature; Gmail delegates (not harassment-focused)	**Limited (creator tools only)**

anti-harassment features (monitoring, blocking, muting, reporting, etc.) their delegates can access. Instagram should extend its “roles” feature from business accounts to all accounts. Facebook should extend admin privileges from pages to profiles.			
Offer all three mechanisms—blocking, muting, and restricting—and apply each of these mechanisms consistently across all different forms of communication, including comments, DMs, tags, and mentions, and across desktop and mobile apps, regardless of device or browser type. Like Twitter and Instagram, Facebook should offer the option to mute by specific kinds of content, such as keywords and emojis. Twitter, Facebook, and Instagram should allow users to mute (rather than just block) DMs. Like Instagram, Twitter and Facebook should allow users to block, mute, or restrict accounts by flagging comments in bulk. Twitter and Facebook should offer	Enhanced blocking/muting features across platforms	X keyword muting; Instagram hidden words; bulk blocking limited	**Partial implementation (varies by platform)**

functionalities akin to Instagram's new restricted feature.			
Regulated services must take steps to ensure that users who have been victims of VAWG, or are exposed to harmful content, are directed to, and are able to access, adequate support. Support can include: (a) information from the online service about actions that can be taken to report, protect from and prevent harm, as well as guidance on the regulated services broader actions to address VAWG, such as outlined in this code of practice (b) signposting and access to websites or helplines dealing with the type of online harm experienced by the user or witnessed by others who may be affected by the content, even if not the designated target; For instance, the National Domestic Abuse Helpline, in England run by Refuge, Tech Safety website, Childline, The Revenge Porn Helpline and Report Harmful Content service and StopNCII. Best practice would follow a "polluter pays" model	Support services for VAWG victims	UK Online Safety Act requirement; platform partnerships with crisis services	**Regulatory requirement (UK), implemented via partnerships**

where financial contributions are made to specialist VAWG organisations providing support (c) reports from children or related to child safety be expedited (d) information from, and contact details for, services providing victim support or mental health support after being exposed to hateful, violent, and harmful materials.			

MORE DESIGN PATTERNS FROM THE TAXONOMY TO GUIDE SAFETY:

Design Pattern	**Example**	**Deployment Stage**
User-controlled visibility	Twitter protected tweets; Instagram private accounts; Facebook friend lists and selective sharing	**Widely deployed**
Privacy tooling	Facebook privacy checkup; Instagram close friends list; Twitter protected tweets	**Widely deployed**

Escalating penalty systems	Twitter/X strike system; YouTube Community Guidelines strikes; Twitch suspension escalation	**Widely deployed (strike system is widely deployed but its difficult for users to know how it's deployed or even know their strikes)**
Default protection settings	Instagram hiding sensitive content by default; Snapchat friend lists hidden for minors	**Implemented (varying by platform)**
Data minimization	Privacy by design principles; limited data retention; aggregate statistics instead of individual tracking	**Regulatory requirement (enforcement varies)**
Appeals processes	YouTube copyright appeals; Meta Oversight Board; general content moderation appeals	**Widely deployed (effectiveness varies)**
Muting content/users	X (Twitter) mute accounts, keywords, hashtags; Instagram mute posts/stories; Facebook "snooze" for 30 days	**Widely deployed**
Data protection integration	GDPR-compliant cookie banners; right to deletion features; data portability tools	**Regulatory requirement (deployed)**

Data Handling	End-to-end encryption (Signal, WhatsApp); minimal data retention policies	Varying deployment
Report tooling	Instagram's multi-step reporting flow; YouTube's video/channel reporting; X bulk tweet reporting	**Widely deployed (quality varies)**
Granular moderation tools and filters	Instagram's "hidden words" feature; TikTok comment filters; YouTube's comment review system	**Increasingly deployed**
Technical safety measures	Two-factor authentication(2FA) (widely available); end-to-end encryption (E2EE) (Signal, WhatsApp); login alerts	Some widely deployed (2FA) but others like E2EE not at all (Signal exemplar)
Automated detection systems	YouTube Content ID; Meta's automated hate speech detection; Twitch AutoMod	Deployed but with limited features - there are automated detection systems but not ones able to flag addresses
Algorithmic detection systems	PhotoDNA for CSAM; Amazon Rekognition	**Widely deployed**
Real-time-Intervention systems	Instagram's "Are you sure you want to post this?" nudge for offensive comments; Twitter/X reply warning system	**Product integration (major platforms) with no cross doxxing alert systems**

Proactive content filtering	Instagram automatic offensive comment hiding; Snapchat nudity detection; spam filters	**Widely deployed**
Flagging	User reports; automated flagging; trusted flagger priority queues	**Widely deployed**
Detection and response	AirTag stalking alerts; suspicious login detection and account lockdown	**Implemented (evolving)**
Shielding content	Block Party lockout folders (discontinued); Reddit Harassment Filter for moderators	**Limited deployment (third-party tools, some platform features)**
One-click documentation	TRFilter documentation (discontinued due to API changes); Instagram report history	**Research/pilot phase (platform resistance)**
Centralized abuse dashboard	YouTube Copyright Match Tool dashboard (exists for IP, not harassment); PEN America recommendation	**Proposed (not implemented for harassment)**
Real-time alerts	Login from new device alerts; AirTag “traveling with you” alerts; unusual activity notifications	**Widely deployed**
Customizable filters	X keyword muting; Instagram hidden words; TikTok keyword filters	**Widely deployed**

Granular controls	Facebook audience selectors per post; Instagram close friends; privacy settings per data type	**Widely deployed**
Revocable permissions	App permission revocation (iOS/ Android); GDPR right to deletion; OAuth token revocation	**Standard (platform dependent)**
Disclosure practices	Privacy policies; transparency reports; algorithmic impact assessments	**Standard (comprehension varies widely)**

MORE SPECIFIC DESIGN COMPONENTS FROM THE TAXONOMY:

Design Components	Deployment Stage
Emergency Hotline	**External partnerships exist; in-platform proposed**
Panic/emergency feature that escalates support and connects users with real time human assistance	**Research/prototype phase**
Ban Members	**Widely deployed**

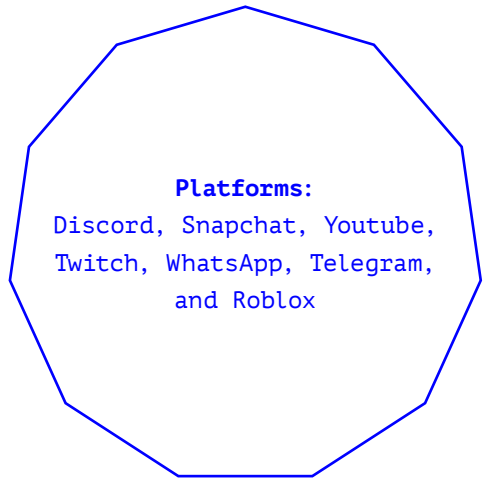
Restrict users	**Implemented**
Timeout Members	**Widely deployed (community platforms)**
Clear out content in bulk	**Regulatory requirement (GDPR), limited bulk tools**
Untag photos	**Widely deployed**
Remove your page likes and interests	**Manual process**
Regular security audits (accountability)	**Common (large platforms), quality varies**
Warnings and flagged content	**Widely deployed**
Capture or export evidence of abuse (screenshots, logs) easily	**Limited implementation (manual screenshots standard)**
“Held for review” queues moderation	**Selectively deployed**
Strike systems to raise accountability for repeat offenders	**Widely deployed (transparency varies)**

Centralized Abuse Dashboard	**Proposed (not implemented for harassment)**
Moderation Dashboard	**Widely deployed (moderator/admin tools)**
Change Nickname	**Widely deployed (community platforms)**
Manage Nicknames	**Widely deployed (community platforms)**
Kick Members	**Widely deployed (community platforms)**

Case Study 5: Children on Gaming & Social Platforms

The victim: 13-year-old Tyler Matthews, facing multi-platform exploitation

Harm: Harassment, mental health decline, anxiety, depression, child endangerment and potential physical safety concerns.



Week 1, harassment starts (Tyler's Perspective): Tyler joins a few Roblox servers where older players make sexual jokes. He joins other Discord servers, where his online "friends" from games start to ask personal questions about his life. Tyler follows different YouTube Gaming Creators, and reads comments that include inappropriate content. He uploads his Minecraft videos to YouTube and starts to receive harassing and mean comments on his videos. Tyler uses Snapchat and Snapchat's Quick Add starts to suggest adult profiles with no mutual connections. Tyler joins Reddit gaming communities that require age verification but he easily bypasses those requirements.

Week 1, harassment starts (Tyler's Parents' Perspective): Tyler's parents notice Tyler seems distracted and secretive about his phone. At first, they aren't worried about his time spent on his computer gaming, because he's been on Minecraft servers with his friends only since he was 11. He and his dad usually talk about Minecraft and Roblox, but Tyler starts to seem cagey about his Roblox usage, and Tyler doesn't mention starting his own YouTube Channel or joining Discord Servers. His parents know he uses Snapchat with friends.

His parents decide to:

- Ask directly but non-accusatory questions like: "Who are you playing with online? Anyone new?"
- His parents decide to proactively set up **TikTok Family Pairing**: this lets them see his activity, who he follows, and set screen time limits. Tyler's DMs are already disabled for under-16 by default
- His parents decide to enable **Roblox Parental Controls**: this requires parent ID verification, shows Tyler's friend list, Tyler's chat settings, and Tyler's spending alerts
- After talking to Tyler about Snapchat and doing some cursory research, his parents turn off **Snapchat Map together** and explain why location sharing is risky.
- His parents decide to limit some of his social media activity, such as how much time he spends on Snapchat

and implement a new family rule of discussing how he's using his social media.

- They decide to review Discord servers with him. During this review, they ask him to show them who's in his gaming servers; this creates a discussion of if Tyler knows the other players and who they are.
- His parents decide to install **Bark**: monitors 30+ platforms for grooming language patterns without reading every message

Month 1, harassment continues to escalate (Tyler's Perspective): Even with Tyler's parents being more involved in his social media usage, Tyler becomes Discord server admin for his friend group without his parents knowledge. Raids from 4chan target his server and shower it with inappropriate and sexualized content. Tyler starts Twitch streaming and receives inappropriate donations with sexual messages. His Instagram Reels and TikTok posts receive comments that start to sexualize his appearance, making him uncomfortable. He doesn't know how to configure safety settings and fears losing followers by being "too strict."

Month 1, harassment continues to escalate (Tyler's Parents' Perspective): Tyler's parents notice he is continuing to withdraw, despite their conversations and best efforts. His parents sit down and express their concerns. His parents say they have noticed some of the messages and comments he's received on Instagram via Instagram Parent Controls. They explain it's okay for him to be uncomfortable with those messages and comments as anyone would feel uncomfortable. They ask him if he's on other websites but he does not share about his own Discord server. For under-18 users, Discord automatically enables explicit image filtering and blocks access to 18+ servers but, Discord cannot be fully monitored without the teen's cooperation[48]. Luckily, TikTok's Family Pairing offers the most comprehensive parental control suite among social platforms by disabling DMs completely for accounts under 16, and so Tyler is not receiving harmful DMs on TikTok[49].

Month 2 and onward, the harassment gets more severe (Tyler's Parents' Perspective): Tyler's parents discover Tyler sent photos to a "girlfriend" who's actually a 40-year-old man who originally messaged him on Roblox and joined Tyler's Discord server. Tyler shared his location Snapchat Map which led to his predator attempting real-world contact. Tyler's parents explain these issues to Tyler, and Tyler agrees to let his parents join his Discord server. Tyler also agrees to not speak to strangers on Roblox. Tyler's parents decide to review Roblox messages with Tyler together. Tyler's parents also filed a report with National Center for Missing and Exploited Children (NCMEC) CyberTipline; this helpline processed 20.5 million reports in 2024[50]. His parents suggest therapy for Tyler and decide to pursue criminal charges and complaints. Tyler's anxiety starts to improve.

48 https://connectsafely.org/discord/
49 https://support.tiktok.com/en/safety-hc/account-and-user-safety/family-pairing
50 https://www.thorn.org/blog/what-the-2024-ncmec-cybertipline-report-says-about-child-safety/

Safety by Design Suggestions to Mitigate Harms: Platforms should better implement mandatory age verification beyond self-declaration. Platforms with known issues with of predatory adult behavior, like Roblox, could implement automatic CSAM (child sexual abuse material) scanning with NCMEC reporting. Platforms could implement better parent notifications for concerning patterns, create more proactive nudging for teens and parents to discuss content more, implement filters or warnings for teens from unknown accounts, and create more enforcement on adult-minor private messaging. Platforms could also allow parents to take over accounts for minors in emergencies and create mandatory safety education before teens access the platform. Education and onboarding for parents and children should also be created for children's accounts.

Intervention	Label	Example	Deployment Stage
Integrate digital literacy and citizenship tools into platforms to educate users on these topics while they engage with the platform.	Integrated digital literacy tools	Facebook/ Instagram digital literacy initiatives; limited integration into core experience	**Limited implementation (often external)**
Regularly prompt users, via nudges and reminders, to review their security and privacy settings and set up the safety modes detailed above. Prompts could proactively encourage users to reconsider including private information that could put them at risk (such as a date of birth or home address).	Regular prompts to review privacy/safety settings	Facebook privacy checkup; Instagram safety reminders	**Implemented (varying frequency/effectiveness)**

Platforms should use nudges to discourage users' attempts to engage in abusive behavior. One way to do this is to use automation to proactively identify content as potentially abusive and nudge users with a warning that their content may violate platform policies and encourage them to revise it before they post.	Nudges to discourage abusive behavior	Instagram "Are you sure?" prompts; X/ Twitter reply warnings	**Implemented (major platforms)**
Using more simple and accessible language throughout the user experience. Ensure that any steps taken to address or help users who are experiencing abuse are clearly communicated and easily accessible.	Simple and accessible language in safety UX	Ongoing effort; varying success; design principle	**Continuous improvement (inconsistent)**
Providing easy navigation and access to safety tools. Ensure that safety tools already available are as easy to find and use as possible.	Easy navigation to safety tools	Safety centers on major platforms; discoverability varies	**Implemented (effectiveness varies)**
Using simple and accessible language throughout the user experience	Simple and accessible language in safety UX	Ongoing effort; varying success; design principle	**Continuous improvement (inconsistent)**

Permanently suspend users if they bully or harass others	Permanent suspension for bullying/ harassment	All platforms have suspension policies; enforcement inconsistency	**Widely deployed (enforcement questioned)**
Require users of these platforms to disclose their real identities	Real identity disclosure requirements	Facebook “real name” policy (controversially abandoned for some); proposed in some jurisdictions	**Controversial (largely not implemented due to safety concerns)**
Offering users the ability to track and manage their reports	Report tracking and management	Instagram report status; YouTube reporting history; proposed comprehensive dashboards	**Partial implementation**
Users must be able to effectively report content that is illegal or harmful to regulated services through clear and transparent flagging mechanisms. Regulated services are obligated to have effective and easy to use reporting functions and must use them to triage content for both human and automated moderation.	Effective and easy-to-use flagging mechanisms	UK Online Safety Act requirement; all platforms have reporting	**Regulatory requirement (UK), widely deployed**
Looking for early warning signs parents often miss	Early warning signs detection for parents	Parental control apps (third-party); platform-level limited	**Third-party tools; limited platform integration**

Age-appropriate access controls or monitoring approaches	Age-appropriate access controls	TikTok DM restrictions for 13-15; Instagram teen accounts with enhanced privacy; COPPA compliance	**Widely deployed (enforcement varies)**
Technical access controls and knowing their limitations	Technical access controls and limitations understanding	Parental control platforms; varying transparency	**Implemented (clarity varies)**
Communication strategies that maintain trust while ensuring safety.	Trust-maintaining communication strategies	Required under UN Guiding Principles; DSA risk assessments; limited public examples	**Regulatory requirement (limited public transparency)**

MORE DESIGN PATTERNS FROM THE TAXONOMY TO GUIDE SAFETY:

Design Pattern	Example	Deployment Stage
Age assurance mechanisms	Age self-declaration (common); age estimation AI (some platforms testing); government ID verification (controversial, limited deployment)	Varying deployment (self-declaration standard)
Age restrictions	TikTok DM restrictions for 13-15; age-gated content on YouTube; COPPA compliance	**Widely deployed (enforcement varies)**

User-controlled visibility	Twitter protected tweets; Instagram private accounts; Facebook friend lists and selective sharing	**Widely deployed**
Appeals processes	YouTube copyright appeals; Meta Oversight Board; general content moderation appeals	**Widely deployed (effectiveness varies)**
Report tooling	Instagram's multi-step reporting flow; YouTube's video/channel reporting; X bulk tweet reporting	**Widely deployed (quality varies)**
Centralized abuse dashboard	YouTube Copyright Match Tool dashboard (exists for IP, not harassment); PEN America recommendation	**Proposed (not implemented for harassment)**
Opt-in defaults	Location sharing off by default; public profiles requiring opt-in; marketing emails requiring consent	**Increasingly common (regulatory pressure)**
Privacy tooling	Facebook privacy checkup; Instagram close friends list; Twitter protected tweets	**Widely deployed**
SOS button with live support	"emergency lockdown mode" or a panic button	**Research/prototype phase**

Informed consent	GDPR consent banners; app permission requests with explanations; Instagram's data usage notifications	**Regulatory requirement (quality varies)**
Flagging	User reports; automated flagging; trusted flagger priority queues	**Widely deployed**
Content hiding mechanisms	Twitter content warnings; Instagram sensitive content blur; Reddit NSFW tags	**Widely deployed**
Revocable permissions	App permission revocation (iOS/ Android); GDPR right to deletion; OAuth token revocation	**Standard (platform dependent)**
Early warning system	Coordinated attack detection; sudden volume spike alerts; behavior pattern analysis	**Research Phase to our knowledge**
Muting content/users	X (Twitter) mute accounts, keywords, hashtags; Instagram mute posts/stories; Facebook "snooze" for 30 days	**Widely deployed**
Real-time-Intervention systems	Instagram's "Are you sure you want to post this?" nudge for offensive comments; Twitter/X reply warning system	**Product integration (major platforms) with no cross doxxing alert systems**

Automated detection systems	YouTube Content ID; Meta's automated hate speech detection; Twitch AutoMod	**Deployed but with limited features - there are automated detection systems but not ones able to flag addresses**
Default protection settings	Instagram hiding sensitive content by default; Snapchat friend lists hidden for minors	**Implemented (varying by platform)**
Data protection integration	GDPR-compliant cookie banners; right to deletion features; data portability tools	**Regulatory requirement (deployed)**
Data Handling	End-to-end encryption (Signal, WhatsApp); minimal data retention policies	**Varying deployment**
Granular moderation tools and filters	Instagram's "hidden words" feature; TikTok comment filters; YouTube's comment review system	**Increasingly deployed**
Technical safety measures	Two-factor authentication(2FA) (widely available); end-to-end encryption (E2EE) (Signal, WhatsApp); login alerts	**Some widely deployed (2FA) but others like E2EE not at all (Signal exemplar)**
Algorithmic detection systems	PhotoDNA for CSAM; Amazon Rekognition	**Widely deployed**

Community driven moderation	Reddit moderator tools; Discord server moderation; Facebook group admin controls	**Widely deployed**
Proactive content filtering	Instagram automatic offensive comment hiding; Snapchat nudity detection; spam filters	**Widely deployed**
Trusted flagger programs	YouTube Trusted Flagger Program; Twitter's former human rights defender program; Global Internet Forum to Counter Terrorism (GIFCT)	**Implemented (varying commitment)**
Detection and response	AirTag stalking alerts; suspicious login detection and account lockdown	**Implemented (evolving)**
Shielding content	Block Party lockout folders (discontinued); Reddit Harassment Filter for moderators	**Limited deployment (third-party tools, some platform features)**
Escalating penalty systems	Twitter/X strike system; YouTube Community Guidelines strikes; Twitch suspension escalation	**Widely deployed (strike system is widely deployed but its difficult for users to know how it's deployed or even know their strikes)**

Rapid response	Twitter's former trusted flagger expedited review; account lockdown during active threats	**Limited implementation (declining)**
Real-time alerts	Login from new device alerts; AirTag "traveling with you" alerts; unusual activity notifications	**Widely deployed**
Language/keyword flagging	Platform-wide profanity filters; user-defined keyword blocks; slur detection	**Widely deployed**
Granular controls	Facebook audience selectors per post; Instagram close friends; privacy settings per data type	**Widely deployed**
Data minimization	Privacy by design principles; limited data retention; aggregate statistics instead of individual tracking	**Regulatory requirement (enforcement varies)**
Purpose limitation	GDPR purpose limitation; advertising preferences tied to specific consents	**Regulatory requirement**
Moderation	Content moderator teams; automated systems; hybrid approaches	**Widely deployed (scale inadequate)**

MORE SPECIFIC DESIGN COMPONENTS FROM THE TAXONOMY:

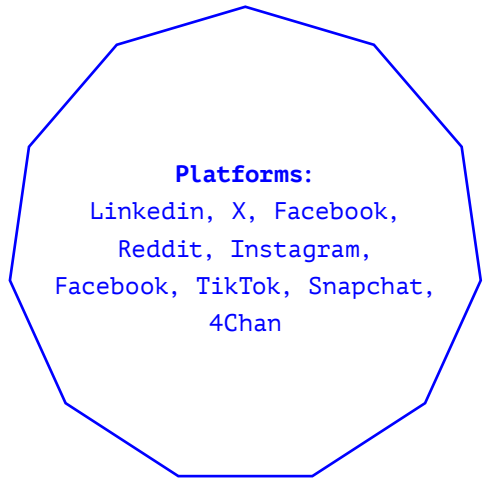
Design Components	Deployment Stage
Capture or export evidence of abuse (screenshots, logs) easily	Limited implementation (manual screenshots standard)
Designate allies who can help monitor, block, report, document abuse on their behalf	Limited (YouTube creators), proposed comprehensively
Panic/emergency feature that escalates support and connects users with real time human assistance	Research/prototype phase
Warnings and flagged content	Widely deployed
Ban Members	Widely deployed
Centralized Abuse Dashboard	Proposed (not implemented for harassment)
Restrict users	Implemented
Emergency Hotline	External partnerships exist; in-platform proposed
Transparent Log	Widely deployed (major platforms)

Regular security audits (accountability)	**Common (large platforms), quality varies**
Default safety settings automatically enabled	Partial implementation (varies by platform and user age)
“Held for review” queues moderation	Selectively deployed
Strike systems to raise accountability for repeat offenders	**Widely deployed (transparency varies)**
Kick Members	**Widely deployed (community platforms)**
Timeout Members	**Widely deployed (community platforms)**
Moderation Dashboard	**Widely deployed (moderator/admin tools)**

Case Study 6: Non-Public User Viral Harassment

The victim: Emma Okonkwo, a regular user's comment goes viral

Harm: Racialized and gender focused harassment, swatting attempts, her workplace is contacted in an attempt to fire her, doxxing attempts, unsolicited visual depiction of intimate parts/image-based sexual abuse



Days 1-3, harassment starts: Emma, a BIPOC woman, comments on a LinkedIn post about workplace diversity. Her comments are screenshotted and posted X where it goes viral with racist and misogynistic responses. Different Facebook profiles share her profile and information with racialized captions about "why diversity hiring fails." The SubReddit's "CringeLinkedIn" mocks Emma's professional photo, engaging in racialized and gendered harassment and comments. Emma receives 500+ LinkedIn connection requests with comments containing harassment from comments about her appearance to violent threats about attacking her. Emma's employer gets tagged on Linkedin asking "is this your DEI hire?", which causes Emma stress. She has to individually block and report each piece of harassment but she can't keep up the pace of harassing comments.

Weeks 1-2, the harassment spreads to other platforms: On Instagram, Emma receives hundreds of DMs with racist slurs and rape threats, prompting her to change her privacy settings and turn off the ability to receive messages from non friends. She receives unsolicited visual depiction of intimate parts/image-based sexual abuse (e.g. dick pics) that flood her Facebook Messenger from international accounts. Facebook Messenger automatically blurs images sent from non-friend accounts but Emma still needs to block and report each account. Emma locks down all of her accounts, and changes her Facebook Messenger account so non friend accounts cannot message her. TikTok videos continue to mock her appearance with millions of views; Emma tries to report each video but there are too many. YouTube Creators create compilation videos titled "Woke Woman Gets Destroyed" to monetize her harassment; Emma also reports these videos but there are too many. Her workplace email gets signed up for thousands of newsletters. Even though she has now made all of her accounts private, there are still screenshots of her, her account and her posts that begin to circulate widely online. Emma is just one of the 25% of Americans who have experienced severe online harassment[51].

Weeks 3 and onward, the harassment continues to escalate: Emma is doxxed and her home address is posted on 4chan with "teach her a lesson" messages, which subtly encourage harassment. Her employer receives bomb threats mentioning her name specifically. A SWATting attempt is made at Emma's home from her address being shared online. Emma takes a leave of absence from work, temporarily relocates her family to a hotel, and tries to figure out her next steps in terms of her harassment. She hires a law firm to pursue legal action, costing her more money. Platforms offer no compensation or direct support to Emma despite her harassment originating from their algorithmic amplification and content.

Safety by Design Suggestions to Mitigate Harms: Platforms could implement emergency support for non-public figures experiencing sudden harassment and one button/panic buttons that can easily configure safety privacy settings or remove identifying information. Platforms could proactively offer nuanced support resources in multiple languages for the different types of harms that users face. Platforms could improve automatic harassment detection especially when content goes viral and is targeting an individual.

51 https://www.pewresearch.org/internet/2021/01/13/the-state-of-online-harassment/

Intervention	Label	Example	Deployment Stage
Institute public-facing community guidelines that address hateful content and harassing behavior, clearly defining the consequences of violations and being clear about substantive changes or exceptions in their guidelines.	Public community guidelines with clear consequences	All major platforms have guidelines; transparency about enforcement varies	**Widely deployed (clarity/ consistency varies)**

Enable users to automatically capture all relevant publicly available data for content that is flagged as abusive via user muting, restricting, blocking, or reporting, as well as proactive content filtering. To ensure that the feature meets evidentiary requirements for legal proceedings, platforms need to work with civil society, law enforcement, and legal experts to create documentation standards in each country and jurisdiction where they operate.	Automatic evidence capture with legal standards	Proposed by PEN America; requires platform-civil society-law enforcement collaboration; Instagram saves limited report history	**Proposed (manual screenshots standard)**
Create a formal, publicly known appeals or escalation process. When users report abusive content that they perceive to be in violation of platform policies and that content is not taken down, they should have accessible avenues to appeal or escalate their case. The appeal or escalation process should be formal, public, integrated into the reporting process, and available to all users who report harassment. When users file an appeal or	Public appeals/ escalation process for dismissed reports	Meta Oversight Board (limited scale); YouTube content appeals; general platform appeals (often opaque); UK Online Safety Act requires appeals processes	**Limited implementation**

escalation, they should be able to amend their case, adding context or additional related abusive content.			
Expand tools and services for targets of harassment, ensuring that social media platforms are easily accessible and effective for those facing or fearing an online attack. Such tools can include the ability to batch report violative content and block multiple perpetrators simultaneously, an abuse-report tracking portal, and real-time rapid-response support.	Comprehensive harassment target tooling for those facing or fearing an online attack.	Proposed by civil society; exists in fragments across platforms (X batch tweet reporting, Instagram report status, but no comprehensive integrated system)	**Fragmented implementation but no comprehensive integrated system**
Clearly articulate policies that prohibit content that harasses or abuses someone on the basis of gender or race.	Prohibit content that harasses or abuses someone on the basis of gender or race.	Flagging systems around harassment policies (hate speech, targeted harassment, etc) although enforcement consistency and comprehensiveness vary significantly	**Widely deployed (effectiveness and enforcement varies on every platform)**
Introduce incident reports: Given the pervasive use of malign creativity, coded language, dogwhistles, and dogpiling, all of which may not be evident in	Incident reports for coordinated abuse	Proposed by researchers; not implemented by platforms	**Proposed (not implemented)**

piece of abusive content, social media platforms should allow users to compile incident reports that provide more context and a more holistic view of the abuse they are experiencing.			
Streamline the reporting process. Platforms should streamline the process by creating a single channel for reporting abusive content or accounts (rather than distinct and divergent channels for harassment, doxing, impersonation, etc.) and requiring as few mandatory steps as possible.	Streamlined single-channel reporting	Most platforms have category selection; not truly single-channel	**Partial (users still navigate categories)**
Establish proactive systems to prevent, manage, and remove online hate speech and harassment against women that are updated regularly to respond to evolving threats, including using AI and engaging networks of fact-checking organizations.	Proactive hate speech prevention systems	Meta's fact-checking partnerships (recently reduced); automated hate speech detection	**Implemented (recent rollbacks noted)**
Create a documentation feature that allows users to quickly and easily record evidence of abuse—capturing screenshots, hyperlinks, and other publicly available data	One-click evidence documentation	Still in research proposal phase - The standard user experience still relies on manual screenshots for complete	**Proposed but not implemented (mostly manual documentation falling short from legal standards)**

automatically or with one click—which is crtical for communicating with employers, engaging with law enforcement, and pursuing legal action.		documentation.	
Social media companies should proactively delete bullying or harassing posts	Proactive deletion of bullying/ harassing posts	Automated systems for clear violations; limited proactive removal for context-dependent harassment	**Limited (mostly reactive reporting-based)**
Permanently suspend users if they bully or harass others	Permanent suspension for bullying/ harassment	All platforms have suspension policies; enforcement inconsistency	**Widely deployed (enforcement quality and consistency across platforms questioned)**
Temporarily suspend users if they bully or harass others	Temporary suspension for bullying/ harassment	Strike/suspension systems on all major platforms	**Widely deployed**
Add bulk reporting. In recognition of the coordinated nature of some harassment campaigns, platforms need to give users the ability to report multiple abusive accounts and content in bulk, to reduce the burden of piecemeal reporting.	Bulk reporting for coordinated harassment	X/Twitter batch tweet reporting	**Limited implementation (Twitter tweets only)**

MORE DESIGN PATTERNS FROM THE TAXONOMY TO GUIDE SAFETY:

Design Pattern	Example	Deployment Stage
SOS button with live support	“emergency lockdown mode” or a panic button	**Research/prototype phase**
Report tooling	Instagram’s multi-step reporting flow; YouTube’s video/channel reporting; X bulk tweet reporting	**Widely deployed (quality varies)**
Privacy tooling	Facebook privacy checkup; Instagram close friends list; Twitter protected tweets	**Widely deployed**
Data Handling	End-to-end encryption (Signal, WhatsApp); minimal data retention policies	**Varying deployment**
Selective sharing	Facebook audience selector; Instagram close friends stories; LinkedIn connection-only posts	**Widely deployed**
User configurable filter tolerance	Twitch AutoMod 4-level severity settings; Reddit Harassment Filter accuracy toggle; YouTube comment strictness levels	**Implemented (selective platforms)**

Rapid response	Twitter's former trusted flagger expedited review; account lockdown during active threats	**Limited implementation (declining)**
Default protection settings	Instagram hiding sensitive content by default; Snapchat friend lists hidden for minors	**Implemented (varying by platform)**
Appeals processes	YouTube copyright appeals; Meta Oversight Board; general content moderation appeals	**Widely deployed (effectiveness varies)**
Moderation	Content moderator teams; automated systems; hybrid approaches	**Widely deployed (scale inadequate)**
Flagging	User reports; automated flagging; trusted flagger priority queues	**Widely deployed**
Muting content/users	X (Twitter) mute accounts, keywords, hashtags; Instagram mute posts/stories; Facebook "snooze" for 30 days	**Widely deployed**
Real-time-Intervention systems	Instagram's "Are you sure you want to post this?" nudge for offensive comments; Twitter/X reply warning system	**Product integration (major platforms) with no cross doxxing alert systems**

Automated detection systems	YouTube Content ID; Meta's automated hate speech detection; Twitch AutoMod	**Deployed but with limited features - there are automated detection systems but not ones able to flag addresses**
Data protection integration	GDPR-compliant cookie banners; right to deletion features; data portability tools	**Regulatory requirement (deployed)**
Algorithmic detection systems	PhotoDNA for CSAM; Amazon Rekognition	**Widely deployed**
Proactive content filtering	Instagram automatic offensive comment hiding; Snapchat nudity detection; spam filters	**Widely deployed**
Truster flagger programs	YouTube Trusted Flagger Program; Twitter's former human rights defender program; Global Internet Forum to Counter Terrorism (GIFCT)	**Implemented (varying commitment)**
Content hiding mechanisms	Twitter content warnings; Instagram sensitive content blur; Reddit NSFW tags	**Widely deployed**
Detection and response	AirTag stalking alerts; suspicious login detection and account lockdown	**Implemented (evolving)**

Shielding content	Block Party lockout folders (discontinued); Reddit Harassment Filter for moderators	**Limited deployment (third-party tools, some platform features)**
User accessible moderation tools	Discord server moderation; Facebook group admin tools; subreddit moderator controls	**Widely deployed**
Comment moderation tools	YouTube comment review folders; Facebook Moderation Assist; Instagram comment controls	**Widely deployed (creators/pages)**
Escalating penalty systems	Twitter/X strike system; YouTube Community Guidelines strikes; Twitch suspension escalation	**Widely deployed (strike system is widely deployed but its difficult for users to know how it's deployed or even know their strikes)**
Centralized abuse dashboard	YouTube Copyright Match Tool dashboard (exists for IP, not harassment); PEN America recommendation	**Proposed (not implemented for harassment)**
Automated content monitoring	Facebook's proactive hate speech detection; CSAM scanning systems	**Widely deployed (varies by harm type)**
Real-time alerts	Login from new device alerts; AirTag "traveling with you" alerts; unusual activity notifications	**Widely deployed**

Language/keyword flagging	Platform-wide profanity filters; user-defined keyword blocks; slur detection	**Widely deployed**
Informed consent	GDPR consent banners; app permission requests with explanations; Instagram's data usage notifications	**Regulatory requirement (quality varies)**
Data minimization	Privacy by design principles; limited data retention; aggregate statistics instead of individual tracking	**Regulatory requirement (enforcement varies)**
User-controlled visibility	Twitter protected tweets; Instagram private accounts; Facebook friend lists and selective sharing	**Widely deployed**
Granular moderation tools and filters	Instagram's "hidden words" feature; TikTok comment filters; YouTube's comment review system	**Increasingly deployed**

MORE SPECIFIC DESIGN COMPONENTS FROM THE TAXONOMY:

Design Components	Deployment Stage
Mass unfriend connections	Limited (mostly manual or third-party)
Capture or export evidence of abuse (screenshots, logs) easily	Limited implementation (manual screenshots standard)
Emergency Hotline	External partnerships exist; in-platform proposed
Panic/emergency feature that escalates support and connects users with real time human assistance	Research/prototype phase
Clear out content in bulk	Regulatory requirement (GDPR), limited bulk tools
Designate allies who can help monitor, block, report, document abuse on their behalf	Limited (YouTube creators), proposed comprehensively
Delete your posts	Widely deployed (varying ease of use)
Archive your posts	Widely deployed
Restrict users	Implemented

Kick Members	**Widely deployed (community platforms)**
Ban Members	**Widely deployed**
Timeout Members	**Widely deployed (community platforms)**
Delete your comments and reactions	**Mostly manual**
Remove your page likes and interests	**Manual process**
Kill switch	**Proposed (limited emergency deactivation exists)**
No metadata collection (private)	**Limited implementation (Signal exemplar)**
Default safety settings automatically enabled	**Partial implementation (varies by platform and user age)**
Strike systems to raise accountability for repeat offenders	**Widely deployed (transparency varies)**
Centralized Abuse Dashboard	**Proposed (not implemented for harassment)**

Change Nickname	**Widely deployed (community platforms)**
Warnings and flagged content	**Widely deployed**
Regular security audits (accountability)	**Common (large platforms), quality varies**

Conclusion, Recommendations and next steps

Safety by Design is an on-going interactive process that requires community and expert input from activists, journalists, policymakers, civil society and academia, and for real transparency and collaboration from platforms to actualize Safety by Design mechanisms and suggestions.

It's imperative to note that many of the interventions, mitigations, and solutions documented in the Safety by Design taxonomy and definition have been demanded by impacted individuals, communities and experts for years. Most, if not all of these suggestions have been threat modeled, analysed, and tested by outside experts and communities and the proposed suggestions often accompany robust information and documentation of how to implement said suggestions, how to augment suggestions for impacted communities, and how to avoid further harm when implementing said suggestions. To fully create more nuanced aspects of safety for users, it's time for platforms to truly implement comprehensive Safety by Design mechanisms that all users, regardless of location and background, can access whenever they choose to.

Safety by Design is understanding that user customization and agency matter, but that platforms also must take on the labor of implementing, maintaining, and innovating Safety by Design mechanisms as well. Platforms have a specific role and labor to play in creating Safety by Design, including: trained trust safety workers and content moderators who receive in-depth and nuanced support from platforms; emergent and interactive trauma informed policy that responds to new forms of harassment; understanding how privacy and security have real roles to play in safety and ensuring opt out over opt in; engaging in transparency with their policies; ensuring legibility, accessibility and usability across design and policy;

actively engaging with communities and outside experts; understanding the true nuances of ensuring privacy and user autonomy while also ensuring users' can take legal action when they must and ensuring users can access the data and documentation they need for said legal action; and the hundreds of other suggestions that our Safety by Taxonomy documents.

There are steps and takeaways that policymakers and platforms can do right now to ensure Safety by Design, including:

1. **Policy makers should use the definition, taxonomy and case studies together** and pass stronger and specific regulation that centers actionable and concrete Safety by Design recommendations for platforms to implement now. This can include requiring specific or standardized safety or security measures that all platforms must implement such as localized helplines, removing hate speech, providing nuanced privacy filters and content filters to allow users to customize and respond to specific harassment they face, blocking, image blurring from content that might be sensitive or explicit in nature, creating nuanced policies in regards to LGBTQIA+ users, and a myriad of other settings and products that could become key requirements.

2. **Policymakers must not limit or target well known safety and security protocols, like encryption, in efforts to protect one community over another.** As the non-profit, Chayn, has written that 'encryption is a feminist issue[52]' which explains that "encryption plays a vital role in safeguarding the rights and safety of women and gender minorities across a range of contexts…we explore three key use cases where encryption serves as a critical safety tool: supporting survivors of intimate partner violence, ensuring secure access to abortion services, and enabling feminist activism." Encryption plays a vital role in protecting victims of intimate partner violence, activists, journalists and many others across the globe from bad actors, authoritarian governments and other harms.

3. **Many of the design interventions documented in the Safety by Design taxonomy have been demanded over and over again for years** as types of design interventions at risk users want. These interventions should be implemented or platforms must provide accountability, transparency and information on what they've tried and why they are not implementing

52 https://blog.chayn.co/encryption-as-a-feminist-issue-a-policy-brief-e6a9366b77ad

these suggestions. Furthermore, there must be ways for communities, activists and civil society to propose interventions even if platforms disagree.

4. **Power must be given to individuals to be able to enact agency and try to customize, and augment their digital experiences to their preferences.** This exact issue is something Bojana Kostic brought up in interviews. Bojana explained, "There must be something in between and many different things in between this kind of a full control that companies have and then full control the individuals have....there is a lot about Safety by Design that I don't know that might be actually really a good set of instruments to literally kind of bring back the power to individuals to control their data so that it actually can work on issues that through the design of these technologies that actually prevent or some way, one way or another, [to] stop harm once it starts."

5. **Demand that platforms invest more money back into trust and safety teams, and safety mitigations.** The gutting of internal teams at platforms who are responsible for trust and safety must be reversed, and platforms must scale up teams and resources to better accommodate, mitigate and tackle harms that users face at scale. This is not a problem that automation can solve alone; people and technology, like automation, must be enacted on platforms to fully address harms that users face and engage in true Safety by Design processes.

6. **Platforms must commit to truly engaging with Safety by Design.** This means understanding that implementation matters, and so does a real commitment to the taxonomy, ongoing policy, and the definition. A platform cannot just try this work once, or treat the definition and taxonomy like a lightweight checklist to attempt to 'safety-wash' their work. Similarly, a platform should not center the needs of one, more privileged group over another, or only commit to tackling one type of problem and not tackle all of the different digital harms that communities face. Safety by Design does mean safety for all, not safety for some.

Bibliography

1. Using 'safety by design' to address online harms John Perrino, July 26, 2022 https://www.brookings.edu/articles/using-safety-by-design-to-address-online-harms/
2. OECD (2024), "Towards digital safety by design for children", OECD Digital Economy Papers, No. 363, OECD Publishing, Paris, https://doi.org/10.1787/c167b650-en
3. https://annecollier.medium.com/hidden-in-plain-sight-safety-innovation-in-live-streamed-video-3a276417ab6d - Yubo + Twitch measures
4. https://www.womensrightsonline.net/evaluation-of-ogbv-tools
5. https://www.womensrightsonline.net/post/malicious-flagging
6. https://www.isdglobal.org/wp-content/uploads/2023/09/Misogynistic-Pathways-to-Radicalisation_Recommended-Measures-for-Platforms-to-Assess-and-Mitigate-Online-Gender-Based-Violence.pdf
7. https://www.isdglobal.org/isd-publications/the-cost-of-doing-politics-gendered-abuse-and-digital-platforms-role-in-undermining-democracy/
8. https://www.isdglobal.org/isd-publications/web-of-hate-a-retrospective-study-of-online-gendered-abuse-in-2022-in-the-united-states/
9. https://igp.sipa.columbia.edu/sites/igp/files/2024-09/IGP_TFGBV_Its_Everyones_Problem_090524.pdf
10. https://commonplace.knowledgefutures.org/pub/trust-through-trickery/release/1
11. https://chayn.notion.site/Orbits-a-global-field-guide-to-advance-intersectional-survivor-centred-and-trauma-informed-interv-8d8dc6a1436543b8b25af63ddafaa409
12. https://cdn.prod.website-files.com/60fdc9111506063bb9fe8e49/64b081438e3221d7ffc92b12_Trauma-informed%20design_%20the%20whitepaper%20by%20Chayn.pdf
13. https://www.forbes.com/sites/gabbyshacknai/2022/04/22/bumble-launches-partnership-with-bloom-to-offer-complimentary-trauma-support-for-sexual-assault-survivors/
14. https://pen.org/report/shouting-into-the-void/
15. https://pen.org/report/no-excuse-for-abuse/
16. https://pen.org/report/treating-online-abuse-like-spam/
17. https://citizenlab.ca/wp-content/uploads/2024/12/Report180-noescape112924.pdf
18. Del-Real, C., De Busser, E., & van den Berg, B. (2025). A systematic literature review of security and privacy by design principles, norms, and strategies for digital technologies. International Review of Law, Computers & Technology, 1–32. https://doi.org/10.1080/13600869.2025.2457227
19. Safety by Design Principles (2021) eSafety Commissioner https://www.esafety.gov.au/industry/safety-by-design#:~:text=It%20emphasises%20accountability%20and%20aims,protect%20those%20most%20at%20risk
20. Safety by Design Overview – a foundational guide to integrating safety at product inception, including principles and implementation tips https://www.esafety.gov.au/industry/safety-by-design
21. Lee A Bygrave Security by Design: Aspirations and Realities in a Regulatory Context lee.bygrave@jus.uio.no pp 126–177 7 June 2022 https://doi.org/10.18261/olr.8.3.2
22. Radanliev, P. Digital security by design. Secur J 37, 1640–1679 (2024). https://doi.org/10.1057/s41284-024-00435-3
23. https://bloom.chayn.co/
24. https://www.blockpartyapp.com/

25. https://socialfactor.com/blog/social-media/a-comprehensive-list-of-social-media-moderation-tools/
26. https://abookapart.com/products/design-for-real-life.html
27. https://abookapart.com/products/design-for-safety
28. https://www.andalsotoo.net/wp-content/uploads/2018/10/Building-Consentful-Tech-Zine-SPREADS.pdf
29. https://www.cca.edu/newsroom/decolonial-school-decolonizing-design/#:~:text=New%20anti%2Doppressive%20futures,mean%20any%20kind%20of%20future.
30. https://www.belfercenter.org/publication/design-margins#:~:text=Design%20From%20the%20Margins%20(DFM,for%20all%20users%20and%20companies.
31. https://designjustice.org/
32. https://www.sciencedirect.com/science/article/pii/S2667096824000703
33. https://academic.oup.com/bjsw/article/54/1/419/7272719
34. https://journals.sagepub.com/doi/10.1177/23333936211028176
35. https://www.amnesty.org/en/what-we-do/technology/online-violence/
36. https://www.gao.gov/products/gao-24-107292
37. https://arxiv.org/abs/2404.16212
38. https://www.sciencedirect.com/science/article/pii/S2772918424000067
39. https://epjdatascience.springeropen.com/articles/10.1140/epjds/s13688-024-00456-3
40. https://www.justsecurity.org/79995/disinformation-radicalization-and-algorithmic-amplification-what-steps-can-congress-take/
41. https://dl.acm.org/doi/fullHtml/10.1145/3614419.3644003
42. https://www.tandfonline.com/doi/full/10.1080/13600869.2025.2457227
43. https://link.springer.com/article/10.1057/s41284-024-00435-3
44. https://www.iiprd.com/algorithmic-amplification-and-defamation-legal-and-ethical-implications-for-digital-platforms/
45. https://www.stockholmresilience.org/news--events/climate-misinformation/chapter-3-how-algorithms-diffuse-and-amplify-misinformation.html
46. http://web.mit.edu/comm-forum/legacy/mit6/papers/Gillespie.pdf
47. https://news.harvard.edu/gazette/story/2019/03/harvard-professor-says-surveillance-capitalism-is-undermining-democracy/
48. https://pen.org/report/shouting-into-the-void/
49. http://web.mit.edu/comm-forum/legacy/mit6/papers/Gillespie.pdf
50. https://journals.sagepub.com/doi/full/10.1177/1461444816661553
51. https://dokumen.pub/the-platform-society-public-values-in-a-connective-world-0190889764-9780190889760.html
52. https://markovicdejan.com/wp-content/uploads/2022/03/Text-1_Sarah-T.-Roberts-Behind-The-Screen_Digital-Humanity.pdf
53. https://www.sfu.ca/~decaste/OISE/page2/files/HarawayCyborg.pdf
54. https://www.hbs.edu/faculty/Pages/item.aspx?num=56791
55. https://chicagounbound.uchicago.edu/cgi/viewcontent.cgi?article=1052&context=uclf
56. https://datasociety.net/library/safiya-umoja-noble-algorithms-of-oppression/
57. https://ieeexplore.ieee.org/document/9592109/
58. https://books.google.com/books/about/Plans_and_Situated_Actions.html?id=AJ_eBJtHxmsC
59. https://www.isdglobal.org/wp-content/uploads/2023/09/Misogynistic-Pathways-to-Radicalisation_Recommended-Measures-for-Platforms-to-Assess-and-Mitigate-Online-Gender-Based-Violence.pdf
60. https://commonplace.knowledgefutures.org/pub/trust-through-trickery/release/1
61. https://www.isdglobal.org/wp-content/uploads/2023/09/Misogynistic-Pathways-to-Radicalisation_Recommended-Measures-for-Platforms-to-Assess-and-Mitigate-Online-Gender-Based-Violence.pdf

62. https://igp.sipa.columbia.edu/sites/igp/files/2024-09/IGP_TFGBV_Its_Everyones_Problem_090524.pdf
63. https://citizenlab.ca/wp-content/uploads/2024/12/Report180-noescape112924.pdf
64. https://www.isdglobal.org/isd-publications/web-of-hate-a-retrospective-study-of-online-gendered-abuse-in-2022-in-the-united-states/
65. https://www.isdglobal.org/isd-publications/the-cost-of-doing-politics-gendered-abuse-and-digital-platforms-role-in-undermining-democracy/
66. https://chayn.notion.site/Orbits-a-global-field-guide-to-advance-intersectional-survivor-centred-and-trauma-informed-interv-8d8dc6a1436543b8b25af63ddafaa409
67. https://cdn.pen.org/wp-content/uploads/2025/05/22182840/Digital-Harassment_-Treating-Online-Abuse-Like-Spam.pdf?_gl=1*1jnfv78*_gcl_au*MTM4MzQ5MzI0Ni4xNzY5NjA5Njkz*_ga*MTMyNTM2NTc5OC4xNzUyMjM5OTAx*_ga_ORQGYGDH22*czE3NzEzMjIzNzUkbzQkZzAkdDE3NzEzMjIzNzUkajYwJGwwJGg0Njc3Mjk1MzE
68. https://www.forbes.com/sites/gabbyshacknai/2022/04/22/bumble-launches-partnership-with-bloom-to-offer-complimentary-trauma-support-for-sexual-assault-survivors/
69. Trust & Safety Professional Association (TSPA): https://www.tspa.org/curriculum/ts-curriculum/safety-by-design/
70. Australian eSafety Commissioner: https://www.esafety.gov.au/industry/safety-by-design
71. World Economic Forum: https://www.weforum.org/projects/safety-by-design-sbd/
72. UK Government Online Safety Priorities: https://www.gov.uk/government/publications/statement-of-strategic-priorities-for-online-safety/final-statement-of-strategic-priorities-for-online-safety
73. Brookings Institution: https://www.brookings.edu/articles/using-safety-by-design-to-address-online-harms/
74. CISA's Secure by Design guidance (October 2023): https://www.cisa.gov/sites/default/files/2023-10/SecureByDesign_1025_508c.pdf - Joint publication with NSA, FBI, and international partners
75. Lawfare article on cybersecurity approaches: https://www.lawfaremedia.org/article/bolt-vs-baked-cybersecurity
76. Infosec Institute on building security in: https://www.infosecinstitute.com/resources/application-security/why-you-should-build-security-into-your-system-rather-than-bolt-it-on/
77. ISACA on Security by Design vs Default (2024): https://www.isaca.org/resources/news-and-trends/isaca-now-blog/2024/security-by-design-or-default
78. NIST CSRC Glossary definition: https://csrc.nist.gov/glossary/term/least_privilege
79. Microsoft Identity Platform guidance: https://learn.microsoft.com/en-us/entra/identity-platform/secure-least-privileged-access
80. CyberArk explanation: https://www.cyberark.com/what-is/least-privilege/
81. BeyondTrust detailed guide: https://www.beyondtrust.com/blog/entry/what-is-least-privilege
82. NIST CSRC definition: https://csrc.nist.gov/glossary/term/defense_in_depth
83. Cloudflare Learning Center: https://www.cloudflare.com/learning/security/glossary/what-is-defense-in-depth/
84. Fortinet resources: https://www.fortinet.com/resources/cyberglossary/defense-in-depth
85. Wikipedia - Defense in depth (computing): https://en.wikipedia.org/wiki/Defense_in_depth_(computing)
86. OWASP Top 10 Proactive Controls - C5: https://top10proactive.owasp.org/the-top-10/c5-secure-by-default/
87. Microsoft Entra Security Defaults: https://learn.microsoft.com/en-us/entra/fundamentals/security-defaults
88. NIST SP 800-53 SA-8(23): https://csf.tools/reference/nist-sp-800-53/r5/sa/sa-8/sa-8-23/
89. Google Cloud secure-by-default organizations: https://cloud.google.com/resource-manager/docs/secure-by-default-organizations
90. Ann Cavoukian's Original Framework: https://iapp.org/media/pdf/resource_center/pbd_implement_7found_principles.pdf

91. Global Privacy and Security by Design Centre: https://gpsbydesigncentre.com/the-seven-foundational-principles/
92. Wikipedia - Privacy by design: https://en.wikipedia.org/wiki/Privacy_by_design
93. Privacy by Design: The 7 Foundational Principles (2011): https://iapp.org/media/pdf/resource_center/pbd_implement_7found_principles.pdf
94. Official EU GDPR: https://gdpr-info.eu/art-25-gdpr/
95. NIST Privacy Framework: https://www.nist.gov/privacy-framework
96. ENISA Privacy by Design guidelines (2014): https://www.enisa.europa.eu/sites/default/files/publications/Privacy%20and%20Data%20Protection%20by%20Design.pdf
97. OWASP Privacy Resources: https://owasp.org/www-project-top-10-privacy-risks/
98. IAPP Privacy by Design Resources Collection: https://web.archive.org/web/20251214131302/https://iapp.org/resources/article/oipc-privacy-by-design-resources/
99. Global Privacy and Security by Design Centre: https://gpsbydesigncentre.com/
100. https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/13817-Delegated-Regulation-on-data-access-provided-for-in-the-Digital-Services-Act_en
101. https://www.techpolicy.press/non-public-data-access-for-researchers-challenges-in-the-draft-delegated-act-of-the-digital-services-act/
102. https://reclaimingautonomyonline.notion.site/Researcher-access-to-platform-data-under-the-DSA-Questions-and-answers-8f7390f3ae6b4aa7ad53d53158ed257c
103. https://dsa40collaboratory.eu/
104. https://policyreview.info/articles/news/can-we-fix-access-to-platform-data
105. https://hateaid.org/en/twitter-landmark-case-antisemitism/

Annex

Terms:

Definition and Examples and Sources:

Safety by Design (SbD): Security by Design is an approach to building systems, software, and products where security considerations are integrated from the very beginning of the development process, rather than being added as an afterthought. SbD is a design paradigm suggesting organizations consider risk of harm to a user or stakeholder group before they occur, rather than after the fact. It involves a systematic and proactive approach to identifying and responding to risks of harm at the earliest stages of product development, from ideation and design through post-launch monitoring. The core principle is that security should be "baked in" rather than "bolted on." This means that during the initial planning, architecture, and design phases, teams proactively identify potential security threats and build protective measures directly into the fundamental structure of what they're creating.

Trust & Safety Professional Association (TSPA): https://www.tspa.org/curriculum/ts-curriculum/safety-by-design/

Australian eSafety Commissioner: https://www.esafety.gov.au/industry/safety-by-design

World Economic Forum: https://www.weforum.org/projects/safety-by-design-sbd/

UK Government Online Safety Priorities: https://www.gov.uk/government/publications/statement-of-strategic-priorities-for-online-safety/final-statement-of-strategic-priorities-for-online-safety

Brookings Institution: https://www.brookings.edu/articles/using-safety-by-design-to-address-online-harms/

CISA's Secure by Design guidance (October 2023): https://www.cisa.gov/sites/default/files/2023-10/SecureByDesign_1025_508c.pdf - Joint publication with NSA, FBI, and international partners.

Lawfare article on cybersecurity approaches: https://www.lawfaremedia.org/article/bolt-vs-baked-cybersecurity

Infosec Institute on building security in: https://www.infosecinstitute.com/resources/application-security/why-you-should-build-security-into-your-system-rather-than-bolt-it-on/

ISACA on Security by Design vs Default (2024): https://www.isaca.org/resources/news-and-trends/isaca-now-blog/2024/security-by-design-or-default

Key aspects of Security by Design:

Proactive Threat Modeling - Teams analyze what could go wrong before building anything, considering how attackers might try to exploit the system and designing defenses accordingly.

CISA's Secure by Design guidance (October 2023): https://www.cisa.gov/sites/default/files/2023-10/SecureByDesign_1025_508c.pdf - Joint publication with NSA, FBI, and international partners

Lawfare article on cybersecurity approaches: https://www.lawfaremedia.org/article/bolt-vs-baked-cybersecurity

Infosec Institute on building security in: https://www.infosecinstitute.com/resources/application-security/why-you-should-build-security-into-your-system-rather-than-bolt-it-on/

ISACA on Security by Design vs Default (2024): https://www.isaca.org/resources/news-and-trends/isaca-now-blog/2024/security-by-design-or-default

Principle of Least Privilege - Users and system components are given only the minimum access rights they need to function, reducing the potential damage if something is compromised.

NIST CSRC Glossary definition: https://csrc.nist.gov/glossary/term/least_privilege

Microsoft Identity Platform guidance: https://learn.microsoft.com/en-us/entra/identity-platform/secure-least-privileged-access

CyberArk explanation: https://www.cyberark.com/what-is/least-privilege/

BeyondTrust detailed guide: https://www.beyondtrust.com/blog/entry/what-is-least-privilege

Defense in Depth - Multiple layers of security controls are built in, so if one layer fails, others still provide protection.

NIST CSRC definition: https://csrc.nist.gov/glossary/term/defense_in_depth

Cloudflare Learning Center: https://www.cloudflare.com/learning/security/glossary/what-is-defense-in-depth/

Fortinet resources: https://www.fortinet.com/resources/cyberglossary/defense-in-depth

Wikipedia - Defense in depth (computing): https://en.wikipedia.org/wiki/Defense_in_depth_(computing)

Secure Defaults - The system is configured to be secure "out of the box" without requiring users to make security-conscious choices.

OWASP Top 10 Proactive Controls - C5: https://top10proactive.owasp.org/the-top-10/c5-secure-by-default/

Microsoft Entra Security Defaults: https://learn.microsoft.com/en-us/entra/fundamentals/security-defaults

NIST SP 800-53 SA-8(23): https://csf.tools/reference/nist-sp-800-53/r5/sa/sa-8/sa-8-23/

Google Cloud secure-by-default organizations: https://cloud.google.com/resource-manager/docs/secure-by-default-organizations

Privacy by Design (PbD): Privacy by Design is a proactive approach to data protection that integrates privacy considerations into the design and development of systems, technologies, and business practices from the outset, rather than as an add-on. Personal data protection is considered throughout the development process, not just added later to comply with regulations. It aims to embed privacy into the core of systems, ensuring that privacy is a fundamental aspect, not an afterthought. This approach helps organizations comply with privacy regulations, build customer trust, and avoid costly data breaches.

Ann Cavoukian's Original Framework: https://iapp.org/media/pdf/resource_center/pbd_implement_7found_principles.pdf

Global Privacy and Security by Design Centre: https://gpsbydesigncentre.com/the-seven-foundational-principles/

Wikipedia - Privacy by design: https://en.wikipedia.org/wiki/Privacy_by_design

IAPP Resources: https://iapp.org/resources/article/oipc-privacy-by-design-resources/

Privacy by Design: The 7 Foundational Principles (2011): https://iapp.org/media/pdf/resource_center/pbd_implement_7found_principles.pdf

Official EU GDPR: https://gdpr-info.eu/art-25-gdpr/

NIST Privacy Framework: https://www.nist.gov/privacy-framework

ENISA Privacy by Design guidelines (2014): https://www.enisa.europa.eu/sites/default/files/publications/Privacy%20and%20Data%20Protection%20by%20Design.pdf

OWASP Privacy Resources: https://owasp.org/www-project-top-10-privacy-risks/

IAPP Privacy by Design Resources Collection: https://iapp.org/resources/article/oipc-privacy-by-design-resources/

Global Privacy and Security by Design Centre: https://gpsbydesigncentre.com/

Validation Workshop

The aim of this workshop was to validate and refine the working definition and draft taxonomy of Safety by Design (SbD) through expert feedback and testing working with real-world case studies. Participants assessed and gave recommendations on how well the taxonomy describes effective safety interventions and where it may need some revision, expansion or reframing.

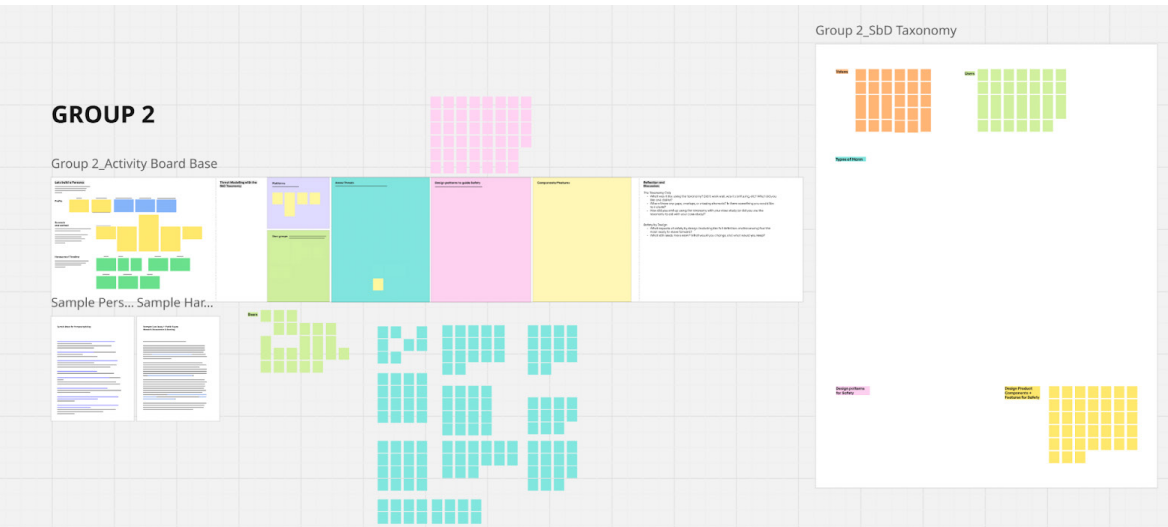
The workshop went as followed:

Part 1: We prepared a presentation showing the main pillars and academic grounding that has been referenced throughout the research to construct the SbD definition and taxonomy.

Part 2: Safety by Design definition was presented to the participants and they provided several key insights to consider to fine tuning the final SbD definition as seen in this report.

Part 3: Participants were introduced to a Miro board that contained the SbD Taxonomy's components and also an activity where they created case studies to test and refine the proposed taxonomy. Participants were divided into breakout groups to create fictional case studies or personas to test the Safety by Design taxonomy to identify potential gaps in the taxonomy and explore how it can be applied to various real-world scenarios.

Part 4: Lastly we had a roundtable discussion where participants provided feedback on their impressions of the exercise, taxonomy in general or any outstanding observations.



Participants were divided into breakout groups to create fictional case studies or personas to test the Safety by Design taxonomy[53].

53 One group discusses creating a case study focused on climate defenders facing multiple challenges, including intimate partner issues and opposition from oil companies.
Another group focused on trans individuals facing hate speech
And lastly around a queer sexworker in Beirut.

Thank You.