EDGELANDS

DESIGNING DIGITAL SAFE SPACES WITH EMPATHY: A TOOLKIT

BACKGROUND

This document represents the culmination of eight weeks of collaborative work by a team of researchers, mentors, and artists, and input from guest speakers, as part of the Edgelands Institute's 'Pop-Down and Beyond' phase.

This is the project output; a toolkit that aims to provide recommendations for the design of digital safe spaces, based on primary source content analysis, case studies, and survey data. We hope that this toolkit may be used by developers, researchers, artists, and online users themselves.

One of the main challenges we identified during our conversations was the lack of inclusion and accessibility of content in languages other than English. Bearing this in mind, we have translated our guide to online safe spaces into Hindi, Russian, German, French, Spanish and Portuguese. We regret that we were not able to translate the toolkit into minority languages, nevertheless we hope that these translations will help share these perspectives and inform action across regions, supporting inclusive dialogue.

ACKNOWLEDGMENTS

This toolkit was developed by Pulkit Mogra, Tatiana Lysova, Lilian Olivia Otero, Catherine Keegan, Nina Martin, Mmabatho Oke, Jessica McClearn, and Giovanna de Custódia, under the guidance of Nina Baranowska, Daniel Odongo, Virgginia Laborão, Vanessa Gathecha, and Laura García Vargas.

We would like to thank everyone who completed our survey and helped shape the recommendations for creating a digital safe space.

Special thanks also go to the guest speakers who joined the Research Sprint's anchoring sessions and shared their insights and inspiration.

The design and visual layout were created by Flavia Lozano and Larissa Oliveira.

EXECUTIVE SUMMARY

This toolkit synthesizes insights from content analysis, case studies, and a qualitative survey to provide recommendations for the design of digital safe spaces that prioritize empathy and specific community needs over extractive platform-centric approaches.

➔ THE PROBLEM

Unlike physical safe spaces, digital environments lack definitive markers of safety. Three key actors shape online safety - platforms, governments, and communities -yet current approaches remain inadequate.

PLATFORMS design for a universalized "average user" (typically cisgender, heterosexual, white, upper-middle-class males from the Global North), rendering marginalized populations' vulnerabilities invisible. Corporate incentives prioritize engagement metrics over user wellbeing

GOVERNMENT REGULATIONS focus narrowly on preventing "tangible harms" (child abuse material, terrorism, fraud), ignoring the subjective, context-specific nature of feeling safe online

COMMUNITIES themselves exercise crucial but fragile safety governance through codes of conduct and volunteer moderation, yet remain structurally subordinate to platform architectures

→ KEY RESEARCH FINDINGS

Our analysis identified four foundational conditions for digital safety:

RELATIONAL CONDITIONS include respecting personal boundaries without requiring confrontation, effective moderation as relational labor (not just rule enforcement), and reducing the emotional cost of participation where users must constantly defend themselves

CULTURAL CONDITIONS encompass social norms of dignity and non-judgment, linguistic inclusivity across languages and communities, and peer networks that provide protective infrastructure for vulnerable groups

PROCEDURAL CONDITIONS involve clear and consistently enforced rules, user autonomy over visibility and data tracking, and legitimate governance structures grounded in lived experiences rather than arbitrary corporate decisions

INFRASTRUCTURAL CONDITIONS address the technical reliability and integrity of platforms, including transparent data practices, effective reporting mechanisms, and accountability when systems fail

These themes are followed by tiered core recommendations - “must-have”, “nice-to-have”, and red flags - for creating safer digital spaces.

➔ CONCLUSION

This toolkit serves as both a practical guide and a call to reimagine digital spaces as community-governed environments where safety is co-created by those who inhabit them, rather than imposed by corporate architectures optimized for engagement over wellbeing. We acknowledge the limitations of our research, including the need for more technical perspectives, broader community involvement, and translation into indigenous languages. Finally, we call for deeper exploration of questions around the topic of online safety.

TABLE OF CONTENTS

I. WHAT ARE DIGITAL SAFE SPACES AND WHAT IS ONLINE SAFETY?

In order to discuss digital safe spaces, we must first delineate what we mean by "digital spaces" and "safety" in online environments.

Despite its ubiquity, the term "digital space" lacks an authoritative, consensus-based definition (Haj-Bolouri et al. 2023). The difficulty in fixing one stems from the continuous reshaping of the purposes, formats, and functions of digital spaces by technological change, and the evolving needs and identities of those who use them (Boyd 2014). Within regulatory frameworks, there has been a tendency to sidestep the term and to instead define the online environment in terms of the platforms that dominate user activity - highlighting the monopolistic power of tech giants (Keskin, 2018; Van Dijk, Poell, and de Wall 2018).

This platform-centric view obscures a fundamental challenge within the concept of "safety". Unlike technical security - which relies on binary, code-based systems - safety is subjective and based on lived experiences and needs of individuals. Different communities will require different levels of protection and security, depending on how at-risk they are. Online safety therefore encompasses not only protection from malware, but also freedom from harm, harassment, and fear (Feldman et al. 2025), as well as the ability to express and present oneself authentically.

Consider how the concept of safe spaces has shifted physically over time. Originating in mid-20th-century liberation movements, initially these were physical sanctuaries such as women's centers or gay bars, where safety was enforced by architecture; a locked door provided protection and a hard boundary. In stark contrast, online safety is a fluid and resource-intensive state that requires active cultivation. This structural ambiguity creates a profound problem: *how can a user verify if a space is safe without a locked door to check?*

➔ WHO CONTROLS THE DOORS?

Online safety currently operates on three dimensions: platforms, governments, and communities.

PLATFORMS

Big Tech companies' de facto control over the online public ecosystem allows them outsized authority to determine what counts as harm, what is permissible, and what forms of risk matter. Platforms dictate the rules and features over which user communities have limited control, such as privacy settings, algorithms, and policies.

Safety is typically modelled on an implicitly universalised "average" user with the risk profile of a cisgender, heterosexual, Caucasian, upper-middle-class male from the Global North. Safety is thereby conceptualised around those who, structurally, face the fewest threats and require the least protection from the vulnerabilities that shape other populations' (online) experiences. Even when attempts are made by research teams inside Big Tech to design for at-risk groups, such learnings are often abstracted to high-level safety needs, given the scale of their platforms.

This leads to a "Wild West" of deregulated or selectively enforced content moderation that can swing dramatically with leadership changes and corporate incentives, and in which content rules can be weakened, enforcement deprioritised, and entire categories of vulnerability rendered invisible. Moreover, there is often a lack of accountability whereby safety decisions happen inside proprietary systems, and are decided by opaque algorithms or internal policy teams with little democratic oversight.

GOVERNMENTS

Top-down regulation - notably the EU's Digital Services Act and the UK's Online Safety Act (OSA) - are self-acclaimed benchmark standards for online safety. However, they are largely limited to the prevention of 'tangible harms' and exposure to a narrow and pre-defined list of online content designated as "illegal" and "harmful", such as child sexual abuse material, terrorism, incitement of violence, and fraud (OSA). Such narrow framings risk flattening the diverse, context-specific ways in which people experience safety online, ignoring that "feeling safe" is subjective and relational - and extends beyond what regulation can easily measure.

COMMUNITIES

User communities themselves exercise a crucial - though under-recognised - form of safety governance. Communities within Discord servers, Facebook groups, Subreddits, etc. routinely establish their codes of conduct, designate moderators, and co-create norms that reflect the values, vulnerabilities, and priorities of their members. Safety then becomes a relational and negotiated practice rather than a top-down imposition.

Communities are often best-positioned to centre marginalised experiences, respond dynamically to emerging harms, and enforce context-sensitive boundaries. However, this form of governance is inherently fragile as it depends on unpaid and unevenly distributed labour, is shaped by internal power dynamics, and remains structurally subordinate to the platform architectures within which it operates. Investigating both the potential and the limits of community-led safety requires closer attention to how users navigate safety in practice.

THE SUBJECTIVITY OF SAFETY

Community members' lack of a shared understanding of what constitutes a safe space presents a key barrier. This subjectivity extends beyond differences in preferences to reflect deeper complexities concerning diverse backgrounds, power dynamics, and technical literacy.

In lieu of locked doors and physical security for online safe spaces, communities must first endeavour to reach a unified understanding of the purpose of the community and what safety is within it. This fosters transparency and psychological safety, whereby members can build trust due to clear boundaries.. Without this, they may suffer from "norm vulnerability" and clashing expectations, which can lead to inadvertent harm; what one person perceives as safety might appear as censorship to another.

The issue of 'context collapse' (Vitak 2012) further exacerbates the subjectivity of assessing safety online. For example, it can be difficult to decipher what is safe banter and what is harassment if one is not rooted in the norms of that given community. Harmful comments may be ignored, or harmless comments may be removed - with either misjudgement impacting users' perception of being safe. Further challenges exist in how a space should be moderated; for example, some neurodivergent users may require explicit written rules while neurotypical users may be comfortable with implied norms.

Ultimately, these challenges are compounded by the restrictions of the platform architecture itself, whereby algorithms are designed to encourage engagement (with content that evokes a strong emotive response achieving that aim) - thereby destabilizing the already-complex feeling of safety that communities endeavour to construct.

➔ THE OBSTACLES POSED BY THE PLATFORMISED MODEL

As discussed, currently online safety is mediated by Big Tech companies and the architecture of their respective platforms. Regarding the specific difficulties in creating online safe spaces that arise, the following are insights from a team member and former online support agent at an international company.

MISALIGNED BUSINESS OBJECTIVES AND A REDUCTIONIST APPROACH TO USERS AND ISSUES

Users are usually viewed as sources of revenue rather than community members, and the business objective is often quantity over quality, e.g., growth of the user base, rather than "customer" retention and satisfaction. Support agents may be expected to "resolve" (often to dismiss) large amounts of user issues in order to satisfy KPIs, rather than actually providing a solution to the problem that the person is facing.

Further, the focus is typically on what is worth fixing or implementing from a profit perspective, thereby addressing the will of the majority or the worst cases of abuse. For example, moderation (both algorithmic and/or human) may be implemented only for dominant languages, permitting unchecked abuse in local language communities (e.g., South Asian languages within an English-speaking server). Safety is therefore defined through corporate interests, such as engagement metrics, rather than user needs or vulnerabilities of certain groups.

LIMITED DISCIPLINARY MEASURES AND CAPACITY TO PROTECT AGAINST MALICIOUS USERS

When banning a user - and communicating with them in general -, agents utilize template messages. There is therefore no constructive conversation or opportunity for repentance, and persistent abusive users may continue to resurface under different IP addresses. Further, the technical skill of certain actors may be greater than that of the company developer team. This can seriously undermine the safety of an online space when there is little capacity to fight against orchestrated malicious acts, which can be extreme as mass-posting pornographic images in spaces designed for young people.

Concerning the protection of children and minors, a user may exhibit predatory behaviour that does not cross the threshold of legal evidence. Therefore, other than "going vigilante" within company hours by monitoring the situation and waiting for it to escalate, agents are limited in their capacity to act.

DISTANCES BETWEEN AGENTS AND COMMUNITIES

Support agents may themselves not be members of the user communities (be that gamers, or cultural groups, etc.), so do not always understand their perspective. There may also be a discrepancy between geographies and cultures of users and support agents, leading to delayed responses due to contrasting timezones and public holidays. Finally, agents wear many hats, working across several products, language communities, and responsibilities, meaning they often do not have the time to provide considered and contextualized support.

In summary, these insights convey two false dichotomies:

01

Firstly, the priority of quantity over quality in terms of user experience belies a dichotomy of maximizing profit versus ensuring safety for all communities. As previously highlighted, platforms generally design for the universal average user whose safety is the least threatened. Though it would be time and resource-intensive to invest in ensuring safety for at-risk populations of all sizes, from female users to minority groups, this would likely bolster customer attraction and retention, thus contributing to the company's success. Further, user safety could itself be a business priority, with tangible aspects included in KPIs; similarly, satisfaction with issue resolutions could be prioritised over their quantity.

02

Secondly, companies' attempts to both streamline their workforce and achieve global reach undermines agents' capacity to care for local contexts. Instead of hiring multilingual and multi-capacity agents, companies could prioritize hiring agents who share a cultural background and timezone with users, who can thus provide dedicated, empathetic, and swift support. As before, this would likely improve customer retention and satisfaction.

➔ THE CASE FOR A COMMUNITY-BASED APPROACH TO DIGITAL SPACES

This toolkit adopts a community-based approach to the design of digital spaces, in conjunction with an expansive understanding of "community". Whether intentionally formed or emerging from contextual or situational factors, a community consists of individuals who share a common identity or set of characteristics that meaningfully shape their (online) experiences.

Defining safety at the level of community permits recognition of differentiated needs, context-specific threats, and the myriad ways in which digital environments can enable or undermine safety. Further, we seek to ensure that definitions of risk, harm, and safety reflect the lived experiences of those who navigate digital spaces, and that those who hold structurally marginalized positions have the opportunity to define what safety is for them.

➔ LEARNING FROM PHYSICAL SAFE SPACES: A CASE STUDY

As safe spaces were initially exclusively physical, we drew inspiration from the policies of physical communities. The following case study is a positive example of a community-cultivated (physical) safe space. Through content analysis, we identified key insights within their community guidelines.

THE "SAFER SPACES POLICY" OF THE REDSTONE SPRING GARDENS ("THE SPRING"), BRISTOL (UK)

Given the nature of the service that The Spring provides - access to an outdoor sauna in a forest setting, where users choose whether to be clothed or not - physical and psychological safety is paramount. To ensure this, The Spring requires users to agree to their community guidelines upon booking a slot.

Within the theme of "Consent, Boundaries, & Body Autonomy", the Spring's guidelines addresses a potential conflict between a users' expectations, as well as the shifting nature of these:

> *Everyone who comes here has different comfort levels, which may shift from moment to moment [...] some people come to connect, others to enjoy quiet time. Please make space for both.*

The acknowledgement of the right to divergent uses of the space is key in creating a foundation for mutual acceptance.

Regarding the navigation of conflict, the guidelines encourage kindness "misunderstandings happen', 'we're all learning'), as well as indicating the presence of an arbiter should something be 'too difficult to handle alone'. These instructions both foster community relations as well as safety, given that staff intervention will occur if necessary.

Within “Privacy & Trust’” the Spring encourages presence ‘with yourself, others and the land’, and forbids photos and recordings, as well as abstaining from sharing any conversations externally without clear consent. These rules encourage conscious individual participation and cultivation, as well as allowing people to share and present themselves authentically.

Finally, the Spring acknowledges the role of each community member: ‘the way you show up helps shape the atmosphere and makes our community thrive’, again reinforcing individual participation. Overall, these guidelines are an excellent example of setting expectations and cultivating a sense of safety from the outset.

SEE ALSO

The “Relaxed Performances” at the Roxy, an event and production venue for independent, contemporary theatre, dance, and performances in Basel, Switzerland.

II. UNDERSTANDING LIVED EXPERIENCES OF USERS OF DIGITAL SPACES: A SURVEY

To understand what online safe spaces mean for diverse communities, five researchers within the team created a qualitative survey to gather data on people's experiences online.

➔ CONTEXT

STUDY DESIGN

We iteratively created questions to explore digital safe spaces from gaps in existing work and knowledge, before sharing it with the wider group for feedback. The questionnaire began with optional demographic questions to help us situate our work in the contexts and identities of respondents, but given the sensitive nature of the work these were optional. These were followed by thematic questions extracting experiences and opinions on topics including lived experiences of digital spaces, possibilities of collaboration between platforms, creators and communities, and the issue of representation in, and access to, digital safe spaces dialogues. The survey was distributed on various online platforms, including the Edgelands Institute social media platforms and within communities that members of the team are affiliated with.

ETHICAL CONSIDERATIONS AND DATA PROTECTION

Due to the lack of an Internal Review Board process, we held multiple meetings concerning the research design and the potential impact on participants. Consequently, all questions were optional to ensure the participants' comfort. Further, this research is embedded in principles of integrity, honesty and transparency in terms of which data we collected and for what purpose. We shared the survey via Notion due to the security of the platform. The data was stored for 60 days to allow for time for analysis being permanently deleted.

SITUATING OUR FINDINGS IN THE DIVERSE IDENTITIES OF PARTICIPANTS

Nine survey responses were collected over 7 days. The average participant age was **39.5 years old**, ranging from **29-57**.

Five identified as women, and four as men.

Seven identified as Caucasian, one as Asian, and one as African.

Three live in Italy, two in Switzerland, and there's one participant each from Canada, England, Germany and Uganda.

English is the primary language for five participants, while French, Italian and Russian are each the primary language for one participant (one participant did not disclose their primary language).

Three participants stated they have a disability, chronic condition or neurodivergence affecting their digital experiences.

Six participants live in urban areas, two in peri-urban and one in a rural area.

Five participants identified as heterosexual, two as bisexual and two as homosexual, with one person preferring not to answer.

Seven participants have personal devices and stable internect access; two have personal devices but sometimes experience connectivity issues.

Seven reported being participants or users of online spaces, with two selecting "other" for this question. Of the seven, two self-identified as digital developers who create safe spaces online.

Relevant communities or identities mentioned include immigrant, indigenous and ethnic minority, LGBTQIA+ and queer, activist and professional, female, hypersensitive, ADHD, and kink.

→ RESULTS AND ANALYSIS

Through reflexive thematic analysis, we identified key themes in the survey data. Overall, we abstract that digital safety is not defined by the tools, filters, or platform-level mechanisms alone, rather by the social and relational environment constructed by a community, within which each interaction takes place - just as in physical spaces. Before safety can be intentionally constructed within a community, certain baseline conditions must be in place. We identified four foundational conditions: relational, cultural, procedural, and infrastructural.

RELATIONAL SAFETY CONDITIONS

Relational safety conditions can be defined as the creation of spaces where individuals can interact openly and authentically without fear of harm.

Relational boundary work: safety is experienced when personal boundaries are respected intuitively without requiring confrontation, explanation, or emotional labour. This sense of "ease of participation" was repeatedly associated with safe spaces.

> *I am on a reddit community of transgender men and we feel safe to share our HRT [hormone replacement therapy] and surgery results to help each other since we know we will not be judged and we will support each other even when we shared issues or results we do not like - survey participant*

Moderation as relational labour: Moderation as relational labour: across responses, effective moderation was described as a relational practice of contacting members, clarifying misunderstandings, de-escalating conflict and modelling fair behaviour. People trust moderators who act with transparency and care.

> *I rely on human support much more [than automated], be it a human moderator of a FB group or a Telegram channel, or simply a discussion on a very supporting online community (for example, a female immigrants group) - survey participant*

Emotional cost of participation: participation becomes labour-intensive when users must constantly defend boundaries, anticipate harassment, or manage interpersonal tensions.

> *In the distant past, I felt neglected as a Facebook user experiencing bereavement, which eventually led me to never use the platform again - survey participant*

CULTURAL SAFETY CONDITIONS

Cultural safety conditions can be understood as online spaces which value and respect the diverse cultural norms and worldviews of communities, leading to environments built on principles of inclusion and empathy, and free from discrimination and harm.

Social norms of dignity: Respectful communication, non-judgment, and an assumption of good faith form a cultural baseline. People describe safety as the ability to exist without having to defend their identity or explain themselves repeatedly. As these form a grounding frame for safety in digital spaces to develop, the following mechanisms are described as those that help construct and maintain safety:

> *It is important for people to understand that even if you interact online there are real people on the other side of the screen and what you say has consequences in the real world - survey participant*

Linguistic inclusivity: Safety is also generated in contexts when users feel included from the point of view of language, e.g. that which is specific to their gender or ethnic minority identity. This also refers to proper translation of all platform and community features, including community guidelines, user agreement, functions, etc..

Local regulation to enhance representation: Community safety can be upholstered with enactment of local (tech) laws. These may include cybersecurity laws, data protection that focus on online community engagement as well as local law enforcement, e.g. protection of minors, or laws against abusive online speech.

The community as a protective infrastructure: For groups exposed to heightened risks (e.g. migrants, queer, and trans communities, neurodivergent users, politically targeted individuals), the community functions as a protective layer. Peer networks provide emotional support, risk assessment, and rapid problem-solving that platforms do not offer.

> *Five years ago, I shared my Instagram account with a person I hardly knew. This person started to stalk me online and [...] in the real life [...], so I sent a desperate post to an immigrant female FB community of the city I am living in. They were very empathetic and expressed their support to me and also provided me with necessary legal info and practical advice, which helped me to scare this person away* - *survey participant*

PROCEDURAL SAFETY CONDITIONS

Procedural safety conditions in digital spaces refer to the specific, actionable steps and ongoing processes that are implemented to protect users. These governance structures can be on different levels including user-to-user or platform-to-user.

Procedural legibility: Clear, transparent and consistently enforced rules form another foundation. Safety depends on understanding how a space works and what to expect from it. When governance systems are arbitrary or opaque, uncertainty can pose harm and hinder any sense of safety from developing.

> *[W]ell-defined guiding principles, openness to discussion [would make online spaces feel more empathetic, community-led, and inclusive]* - *survey participant*

Autonomy of visibility: Respondents consistently linked safety to having control over what is seen, tracked, or collected about them. This extends into questions of agency and power: being able to refuse surveillance or obscured forms of tracking is experienced as a prerequisite for dignity and participation.

> *[Safe space means t]o be acknowledged of who/what is tracking my activity [and t]o have the possibility to decline tracking/surveillance* - *survey participant*

Trust-generating governance: People experience safety when governance structures feel legitimate. This includes predictable rule enforcement, opportunities to participate in decision-making, and policies grounded in lived experiences. While respondents highlighted mechanisms of producing safety, they also described how safety collapses through insidious platform practices.

> *'I feel unsafe when companies often share the data that we give them. Like Facebook showing me ads about what I speak -* survey participant

INFRASTRUCTURAL SAFETY CONDITIONS

Infrastructural safety conditions refers to the reliability and integrity of the digital infrastructure of online spaces.

Epistemic opacity: Unclear rules, hidden forms of data collection, inconsistent moderation, or unexplained account actions generate anticipatory fear and self-censorship. A lack of clarity becomes its own form of harm.

Symbolic moderation failures: Many respondents described reporting mechanisms or community guidelines that may exist but may not function. These failures erode trust because they create a false impression of care or accountability.

Governance as domination: Communities become unsafe when moderators or admins misuse power, enforce discriminatory norms, or act unpredictably. Similarly, platform-level inaction in the face of harm produces a sense of structural abandonment.

> *[I]n our immigrant community there was a rule that all the communication should be "ecological", meaning respectful and not harming others. However, the admin treated this rule in a way that all the communication with them should be this way, while their communication with others overstepped others' boundaries* - *survey participant*

III. HOW TO IDENTIFY, CREATE, AND MANAGE ONLINE SAFE SPACES: A GUIDE

The following is a guide to online safe spaces, based on our research.

➔ OBLIGATORY ELEMENTS

POLICY AND DOCUMENTATION

Clear and well-formulated documentation plays a crucial role in the feeling of safety in digital spaces. Safety increases when rules are clear, explicit, accessible, and visible prior to joining a group. This transparency allows individuals to make informed decisions about whether a digital space aligns with their values and needs and, consequently, whether they wish to join it or not.

Rules should clearly articulate acceptable behaviours, communication norms, and corresponding consequences. This contributes to the feeling of safety by reducing uncertainty and understanding that harmful behaviour will be discouraged. Additionally, this articulation should be in accessible language to all potential community members.

Good practice community rules explicitly address and prohibit harassment, bullying, hate speech, extremist content, and forms of identity abuse (e.g., impersonation, using AI to create inappropriate images of someone without their consent). If rules are somewhat ambiguous, there should be space for their discussion, revision, and reformulation. Replacing broad or abstract language, e.g. “ecological” or “positive” communication, with explicit expectations about respect, non-judgement, and transparency helps make norms more understandable and enforceable.

Community guidelines must be cultivated with input from the communities themselves. Furthermore, they should be continuously and iteratively revisited in line with community members' experiences on the ground.

Data governance should be clearly stated, clarifying how personal data is collected, used, stored, and protected. Ethical data practices, privacy-by-design principles, and transparent and accessible communication about data protection measures are essential to feeling safe in digital spaces. If a digital tool is multilingual, it is vital that all the elements, including those pertaining to data protection, are translated in all the languages.

Finally, rules are effective only when they are applied consistently and equally to all, including moderators and admins. When rules enforcement is uneven or perceived as arbitrary, trust quickly erodes and feelings of exclusion or vulnerability increase.

TECHNICAL ELEMENTS

Robust and up-to-date safety tools are vital components of digital safety, but simultaneously they should be privacy-respecting, context-aware, and non-punitive. Key technical elements that contribute to a sense of safety include content filters, automated removal of disturbing or illegal content, and reporting mechanisms. Encryption and anonymisation mechanisms, data leak prevention, up-to-date security measures, such as two factor authorisation or protections against data leaks, are also crucial in producing the feeling of safety in digital spaces.

However, automated moderation often overreaches,

removing harmless content due to missing contextualisation or too strict rules. Such overreach can limit participation, silence legitimate expression, and undermine confidence in the platform. Automated moderation should therefore work in tandem with human oversight. Additionally, it should be possible to contact the platform with an opportunity to challenge a too strict automated moderation by providing contextualisation and explanation for the content.

Technical protections should also extend to prevent community abuse, private data abuse, and hostile infiltration of sensitive or vulnerable communities. Safety features are more effective when they proactively limit these risks - rather than reactively applied once harm has occurred.

Users should have control over their own privacy on a platform. Platforms should provide tools to manage the visibility of their profiles, selectively disclose personal information, and choose between public and private modes of participation. These features grant users a greater level of control over their exposure, according to their needs and risks.

Real-name policies may be harmful when anonymity and privacy are important for safety. For many users, especially those from marginalised communities or politically sensitive contexts, anonymity is not a preference but a vital protective measure.

Finally, platforms should actively and pre-emptively

"bake-in" adjustments and options for individuals or communities with specific needs, thereby making accommodation the norm rather than a luxury or something that one has to argue for.

COMMUNITY ELEMENTS

In the absence of physical boundaries, users rely on a set of informal yet powerful digital cues to assess whether a space is safe. They may scan for visual markers - effectively reading the platform's "body language"; pronouns in bios, inclusionary symbols, content warnings, or Codes of Conduct pinned to the top of a feed function as immediate signals that boundaries exist and that the space is being intentionally cultivated.

These visual signals are reinforced by linguistic cues that shape the interpersonal atmosphere of a community. Safe spaces tend to adopt inclusive "we" language rather than adversarial "us versus them" rhetoric, and frequently use tone indicators to minimise ambiguity. Inclusivity, including in language, and recognition of diversity are important elements of community guidelines, rules, and codes of conduct. Such practices are particularly important for neurodivergent users, for whom explicit communication reduces anxiety and misinterpretation.

To avoid unwanted infiltration, communities may introduce a basic verification process for new members. For example, there could be an onboarding step requiring new members to acknowledge the

purpose, values, and boundaries of the space before participating and gaining full access..

Ultimately, however, the most decisive cue is behavioural. Community members observe how leadership and moderation operate in practice, judging safety by the speed, consistency, and transparency of responses to harm and violation of the community rules. When hate speech is allowed to linger, or enforcement appears arbitrary, written rules lose credibility and the sense of safety rapidly dissolves.

In case of conflict within a community, rather than relying solely and immediately on reporting and banning tools, platforms should encourage members to first resolve conflicts through kindness and constructive conversation, thus reaffirming the role of the individual in the community. However, disciplinary measures should come swiftly where resolution is not possible.

➔ OUR RECOMMENDED CHECK LIST

"MUST-HAVE" / MINIMUM REQUIREMENTS

- ◯ Clear, explicit, and accessible community guidelines and rules, and intervention when these are broken
- ◯ Strong data protection policy
- ◯ Restricted screenshot function within private communities
- ◯ Reliable reporting tools for harmful content
- ◯ Optional anonymity
- ◯ Oversight for moderators and admins
- ◯ Human moderators (who have access to psychological support and resources if engaging with triggering content)
- ◯ Inclusive language
- ◯ Reactive and proactive measures against harmful speech or activity
- ◯ For multilingual platforms, all elements properly translated into all languages (especially local ones), including user agreements, codes of conduct, etc.
- ◯ Dedicated in-house cybersecurity capacity, especially for at-risk communities

"NICE TO HAVE"

- ◯ Vetting during onboarding process
- ◯ Privacy and protective measures "baked-in" to the platform infrastructure
- ◯ Educational resources on behaviour and etiquette in digital spaces
- ◯ Participatory co-design that incorporates community needs and preferences
- ◯ Support agents / moderators who are members of the user community or at least have knowledge of cultural context, and who have the time and emotional capacity to provide considered support
- ◯ Real time and direct human support, e.g., a helpline available across timezones
- ◯ Dedicated algorithmic tools to enforce specific local norms, including in minority or local languages
- ◯ Structured digital hygiene, e.g., calm or night-time in which moderators and admins may have such tools as limiting messaging to one per minute

RED FLAGS

- ◯ Inconsistency in rules application, such as censorship of diverse non-harmful ideas
- ◯ Persistent lack of response to harm, or harm permitted as "free speech"
- ◯ Obligatory exposure of real identity
- ◯ "Performative" safety, e.g., copying and pasting community rules by default
- ◯ Unprotected technical infrastructure, e.g., exposure to infiltration and data leaks
- ◯ Reliance on algorithms or generic terms of service to handle interpersonal conflict

IV. REFLECTIONS AND LIMITATIONS

Regarding our shared experience of the research sprint that led to this toolkit, we acknowledge here the strengths and limitations that influenced our work together.

The group's interdisciplinary and international composition proved to be valuable. The combination of academic and practical backgrounds, along with a variety of cultural viewpoints, enriched our discussions and the case studies we examined. Although the absence of more technical perspectives left a gap, our collective effort still provided a shared understanding of what safe spaces can mean to different communities.

Designing with empathy became a key part of our process. While empathy can guide us towards more thoughtful solutions, however, we learned that it cannot deliver a universal one. Due to differences in context, language, and access, our toolkit will inevitably serve some communities better than others. The inability to translate the toolkit into indigenous languages, for example, highlights the challenges posed by resource constraints, language preservation, and accessibility. These factors, combined with limited time, shaped what we were able to accomplish. Looking ahead, involving communities more directly in shaping and responding to a further, more expansive toolkit would strengthen its relevance and impact.

Exploring safety in both physical and online environments helped us to understand how different settings influence engagement. At the same time,

the heavy reliance on screens during the sessions emphasised the need for more interactive or embodied activities, which could have improved the experience.

Finally, this research sprint was in of itself an example of a co-creation of an online safe space, in which it was first necessary to establish a foundation by agreeing upon a group definition of community and safety, among other terms.

V. FUTURE CONVERSATIONS AND RESEARCH AVENUES

This toolkit scratches just the surface of the broader and complex challenge of creating safe spaces. We view our survey as a starting point for further investigation through semi-structured interviews or ethnographic work, as there are many open questions that invite deeper exploration and continued dialogue.

A central issue is governance: when community values conflict with platform policies, who should ultimately have the authority to decide what is permissible? This question is made even more pressing by the immense influence of platform monopolies, which control the online information economy and shape power structures, inclusivity, and demographic representation.

We need to better understand how information integrity can serve as a foundation for safer online environments, and what kinds of business or incentive models might empower communities rather than merely extracting value from them. Similarly, envisioning cooperative, co-owned, or otherwise more democratic digital spaces raises important questions about ownership and agency in online life.

Drawing lessons from subversive movements that have successfully created alternative spaces, we may explore how vulnerable groups can disengage from harmful platforms. In tandem, we must consider how governments or infiltrators can repurpose platforms for their own objectives, and what safeguards are necessary to protect communities from such interference.

We note that caution is required in repressive contexts where law may create the conditions for involuntary state-private platform collaboration. Stronger civil society engagement in digital rights is imperative in such cases.

In this toolkit we highlight the need for human moderators who have access to psychological support and resources. This points to the dire need for the implementation of comprehensive and standardized provision of psychological support for moderators.

Finally, we must consider how the knowledge contained within this toolkit can be utilised to create a broader social impact. What kinds of incentives or pressures, for example, might encourage platforms to adopt it as part of their corporate social responsibility efforts, thereby extending its reach and influence?

LANDSLANDS
EDGEEDGE
LANDSLANDS

REFERENCES

Boyd, D. (2014). It's Complicated: The Social Lives of Networked Teens. Yale University Press.

Haj-Bolouri, A., et al. (2023). Typification and Characteristics of Digital Safe Spaces. Conference: 57th Hawaii International Conference on System Sciences. Oxford University Press.

Keskin, Batuhan. (2018). Van Dijk, Poell, and de Wall, The Platform Society: Public Values in a Connective World (2018).

The Redstone Spring Gardens Safer Spaces Policy. Retrieved from: https://mailchi.mp/8d4f7d1752a9/safer-spaces-policy

Vitak, J. (2012). The impact of context collapse and privacy on social network site disclosures. Journal of Broadcasting & Electronic Media, 56(4), 451-470.

Zeena Feldman, Kate Miltner, Zoë Glatt, Sophie Bishop & Ysabel Gerrard. (2025) Algorithms for her? 2: feminist approaches to digital infrastructures, cultures and economies. Journal of Gender Studies 34:8, pages 1107-1117.

EDGELANDS

edgelands.institute