EDGELANDS

EMPATHIEBASIERTES DESIGN SICHERER DIGITALER RÄUME: EIN TOOLKIT

HINTERGRUND

Dieses Dokument ist das Ergebnis einer achtwöchigen Zusammenarbeit zwischen einem Team aus Forschenden, Mentor*innen und Künstler*innen. Zusätzlich gab es Beiträge von Gastredner*innen im Rahmen der Phase "Pop-Down and Beyond" des Edgeland Institutes.

Das Ergebnis dieses Projekts ist ein Toolkit, das Empfehlungen für die Gestaltung sicherer digitaler Räume auf der Grundlage von Fallstudien, Umfragedaten und Quelltextanalysen enthält. Wir hoffen, dass dieses Toolkit von Programmierer*innen, Forschenden, Künstler*innen, Designer*innen und Online-Nutzer*innen selbst genutzt werden kann.

Um mehr über den komplexen Kontext der digitalen Sicherheit, die beteiligten Akteur*innen und die Definitionen von „digitalen sicheren Räumen" und „Online-Sicherheit" zu erfahren, gibt es hier die erweiterte Version dieses Toolkits (nur auf Englisch).

DANKSAGUNG

Dieses Toolkit wurde von Pulkit Mogra, Tatiana Lysova, Lilian Olivia Otero, Catherine Keegan, Nina Martin, Mmabatho Oke, Jessica McClearn und Giovanna da Custódia unter der Begleitung von Nina Baranowska, Daniel Odongo, Virgginia Laborão, Vanessa Gathecha und Laura García Vargas entwickelt.

Wir danken allen, die an unserer Umfrage teilgenommen und zur Entwicklung der Empfehlungen für die Schaffung eines sicheren digitalen Raums beigetragen haben.

Ein besonderer Dank gilt auch den Gastredner*innen, die an den kollektiven Treffen des Research Sprint teilgenommen und ihre Erkenntnisse und Inspirationen mit uns geteilt haben.

Das Design und das visuelle Layout wurden von Flavia Lozano und Larissa Oliveira gestaltet.

ZUSAMMENFASSUNG

Dieses Toolkit fasst Erkenntnisse aus Inhaltsanalysen, Fallstudien und einer qualitativen Umfrage zusammen, um Empfehlungen für die Gestaltung sicherer digitaler Räume zu geben, die empathische Ansätze und spezifische Bedarfe der jeweiligen Community gegenüber extraktiven, plattformzentrierten Ansätzen priorisieren.

➔ DAS PROBLEM

Im Gegensatz zu physischen Schutzräumen gibt es in digitalen Umgebungen keine eindeutigen Sicherheitsmerkmale. Drei wichtige Parameter prägen die Online-Sicherheit - Plattformen, Regierungen und Communities - doch die derzeitigen verwendeten Ansätze sind nach wie vor unzureichend.

PLATTFORMEN ssind für eine*n universelle*n "Durchschnittsnutzer*in" konzipiert (in der Regel cis-geschlechtliche, hetersexuelle, weiße Männer der oberen Mittelschicht aus dem Globalen Norden), wodurch die Vulnerabilität marginalisierter Bevölkerungsgruppen teils unsichtbar wird. Profit- und Unternehmensanreize priorisieren Engagement-Kennzahlen der Nutzer*innen gegenüber ihrem Wohlbefinden.

STAATLICHE VORSCHRIFTEN konzentrieren sich auf die Verhinderung bestimmter Verstöße und Gewalttaten (Kindesmissbrauchsmaterial, Terrorismus, Betrug) und ignorieren dabei die subjektive, kontextspezifische Natur des Sicherheitsgeühls im Internet.

Die **COMMUNITIES** selbst üben durch Verhaltensregeln und freiwillige Moderation eine wichtige, aber fragile Aufgabe des Beitrags zum Sicherheitsgefühl aus, bleiben jedoch strukturell der Plattformarchitektur untergeordnet.

→ SCHLÜSSELERGEBNISSE DER FORSCHUNG

Unsere Analyse hat vier grundlegende Voraussetzungen für digitale Sicherheit identifiziert:

ZWISCHENMENSCHLICHE BEDINGUNGEN umfassen die Achtung persönlicher Grenzen ohne den Bedarf von Konfrontation, effektive Moderation als Beziehungsarbeit und nicht nur zur Durchsetzung von Regeln und die Verringerung der emotionalen Last der Teilnahme, bei der Nutzer*innen sich ständig verteidigen müssen.

KULTURELLE BEDINGUNGEN umfassen soziale Normen der Würde und Nicht-Beurteilung, sprachliche Inklusivität über Sprachen und Gemeinschaften hinweg und Peer-Netzwerke, die eine schützende Infrastruktur für vulnerable Gruppen bieten.

Zu den **VERFAHRENSTECHNISCHEN BEDINGUNGEN** gehören klare und konsequent durchgesetzte Regeln, die Autonomie der Nutzer*innen in Bezug auf Sichtbarkeit und Datenverfolgung sowie legitime Governance-Strukturen, die auf gelebten Erfahrungen und nicht auf willkürlichen Entscheidung von Unternehmensgremien beruhen.

Die **INFRASTRUKTURELLEN BEDINGUNGEN** beziehen sich auf die technische Zuverlässigkeit und Integrität von Plattformen, einschließlich transparenter Datennutzungpraktiken, wirksamer Meldemechanismen und Rechenschaftspflicht bei Systemausfällen.

Auf diese vier Grundthemen folgen dann abgestufte Kernempfehlungen - “must-haves”, “nice-to-haves” - und Warnsignale - “red flags” - für die Schaffung sicherer digitaler Räume.

➔ SCHLUSSFOLGERUNG

Dieses Tooklkit dient sowohl als praktischer Leitfaden als auch als Aufruf, digitale Räume als gemeinschaftlich verwaltete Räume neu zu konzipieren, in denen Sicherheit von Nutzer*innen gemeinsam geschaffen wird und nicht von Unternehmensstrukturen auferlegt wird, die eher auf Engagement als auf Wohlbefinden ausgerichtet sind. Wir sind uns der Grenzen unserer Forschung bewusst, darunter der Notwendigkeit weiterer (technischer) Perspektiven, einer breiteren Einbindung der Communities und der Übersetzung in diverse indigene Sprachen. Abschließend fordern wir eine eingehendere Untersuchung von Fragen rund um das Thema Online-Sicherheit.

WIE MAN SICHERE DIGITALE RÄUME IDENTIFIZIERT, ERSCHAFFT UND VERWALTET: EIN LEITFADEN

Hier folgt ein Leitfaden für sichere digitale Räume, der auf unserer Forschung beruht.

➔ ZWINGENDE BAUSTEINE

RICHTLINIEN UND DOKUMENTATION

UEine klare und gut formulierte Dokumentation spielt eine entscheidende Rolle für das Gefühl von Sicherheit in digitalen Räumen. Dieses Gefühl erhöht sich, wenn die Regeln klar, eindeutig, zugänglich und sichtbar sind, bevor man einer Online-Gemeinschaft beitritt. Diese Transparenz ermöglicht es Einzelpersonen, fundierte Entscheidungen darüber zu treffen, ob ein digitaler Raum ihren Werten und Bedürfnissen entspricht und ob sie ihm beitreten möchten oder nicht.

Regeln sollten angebrachte und erwartete Verhaltensweisen, Kommunikationsnormen und entsprechende Verhaltenskonsequenzen klar formulieren. Dies trägt zum Sicherheitsgefühl bei, indem es Unsicherheiten verringert und deutlich macht, dass störendes und missbräuchliches Verhalten unterbunden wird. Darüber hinaus sollte diese Formulierung in einer für alle (potenziellen) Community-Mitglieder verständlichen Sprache erfolgen..

Bewährte Community-Regeln befassen sich ausdrücklich mit Belästigung, Mobbing, Hassreden, extremistischen Inhalten und Formen des Identitätsmissbrauchs (z.B. Identitätsdiebstahl, Verwendung von KI zur Erstellung unangemessener Bilder von Personen ohne deren Zustimmung) und verbieten diese. Wenn Regeln etwas unklar sind, sollte Raum für deren Diskussion, Überarbeitung und Neufassung gegeben sein. Durch das

Ersetzen von weit gefassten oder abstrajten Begriffen wie “ökologische” oder “positive” Kommunikation durch explizite Erwartungen in Bezug auf Respekt, Nicht-urteilen und Transparenz werden Normen verständlicher und besser durchsetzbar.

Community-Richtlinien müssen unter Einbeziehung der Communites selbst entwickelt werden. Darüber hinaus sollten sie kontinuierlich und wiederholt überarbeitet werden, um die Erfahrungen der Community-Mitglieder in den Räumen einzubinden.

Die Datenverwaltung sollte klar formuliert sein und erläutern, wie personenbezogene Daten erfasst, verwendet, gespeichert und geschützt werden. Ethische Datenpraktiken, Grundsätze des eingebauten Datenschutzes und eine transparente und zugängliche Kommunikation über Datenschutzmaßnahmen sind unerlässlich, um sich im digital Raum sicher fühlen zu können. Wenn ein digitales Tool mehrsprachig ist, ist es wichtig, dass alle Elemente, einschließlich derjenigen, die den Datenschutz betreffen, in alle Sprachen übersetzt werden.

Letztendlich sind Regeln nur dann wirksam, wenn sie konsequent und gleichmäßig auf alle angewendet werden, einschließlich Moderator*innen und Administrator*innen. Wenn die Durchsetzung von Regeln ungleichmäßig ist oder als willkürlich empfunden wird, schwindet das Vertrauen schnell und das Gefühl der Ausgrenzung oder Vulnerabilität nimmt zu.

TECHNISCHE ELEMENTE

Robuste und aktualisierte Sicherheitstools sind wichtige Bestandteile der digitalen Sicherheit, sollten jedoch gleichzeitig die Privatsphäre respektieren,

kontextsensibel und nicht auf einem Strafansatz basieren. Zu den wichtigsten technischen Elementen, die zu einem Gefühl der Sicherheit beitragen, gehören u.a. Inhaltsfilter, die automatische Entfernung störender oder illegaler Inhalte und Meldemechanismen. Verschlüsselungs- und Anonymisierungsmechanismen, Maßnahmen zur Verhinderung von Datenlecks und aktuelle Sicherheitsmaßnahmen wie Zwei-Faktor-Authentifizierung oder Schutz vor Datenlecks sind ebenfalls entscheidend für die Schaffung eines Gefühls der Sicherheit im digitalen Raum.

Allerdings geht die automatisierte Moderation oft zu weit und entfernt harmlose Inhalte aufgrund fehlender Kontextualisierung oder zu strenger Regeln. Ein solches Übermaß kann die Beteiligung einschränken, legitime Meinungsäußerungen unterbinden und das Vertrauen in die Plattform untergraben. Die automatisierte Moderation sollte daher mit einer menschlichen Aufsicht einhergehen. Darüber hinaus sollte es möglich sein, die Plattform zu kontaktieren, um eine zu strenge automatisierte Moderation anzufechten, indem man Kontextualisierung und Erklärungen für den Inhalt liefert.

Technische Schutzmaßnahmen sollten auch auf die Verhinderung von Missbrauch in Communities, Missbrauch privater Daten und feindliche Infiltration vulnerabler Communities ausgedehnt werden. Sicherheitsfunktionen sind wirksamer, wenn sie diese Risiken proaktiv begrenzen, anstatt erst reaktiv eingesetzt zu werden, wenn bereits Schaden entstanden ist.

Nutzer*innen sollten die Kontrolle über ihre Privatsphäre auf einer Plattform haben. Plattformen sollten Tools bereitstellen, mit denen sie die Sichtbarkeit ihrer Profile verwalten, personenbezogene Daten selektiv offenlegen und zwischen öffentlichen und privaten Teilnahmemodi

wählen können. Diese Funktionen gewähren Nutzer*innen ein höheres Maß an Kontrolle über ihre Sichtbarkeit, entsprechend ihren Bedürfnissen und Risiken.

Richtlinien für Klarnamen können schädlich sein, wenn Anonymität und Datenschutz für die Sicherheit wichtig sind. Für viele Nutzer*innen, insbesondere aus marginalisierten Communities oder politisch sensiblen Kontexten, ist Anonymität keine Präferenz, sondern eine wichtige Schutzmaßnahme.

Abschließend sollten Plattformen aktiv und präventiv Anpassungen und Optionen für Personen oder Communities mit besonderen Bedürfnissen einbauen, damit diese Anpassungen zur Norm werden und nicht länger ein Luxus oder etwas sind, um das Nutzer*innen kämpfen müssen.

COMMUNITY-ELEMENTE

Da es keine physischen Grenzen gibt, verlassen sich Nutzer*innen auf eine Reihe informeller, aber aussagekräftiger digitaler Hinweise, um zu beurteilen, ob ein Raum sicher ist. Sie suchen möglicherweise nach visuellen Markern - sie lesen also gewissermaßen die "Körpersprache" der Plattform. Pronomen in Biografien, Symbole für Inklusion, Inhaltswarnungen oder Verhaltensregeln, die oben in einem Feed angeheftet sind, dienen als unmittelbare Signale dafür, dass Grenzen existieren und dass der Raum bewusst gepflegt wird.

Diese visuellen Signale werden durch sprachliche Hinweise verstärkt, die die zwischenmenschliche Atmosphäre einer Community prägen. Sichere Räume neigen dazu, eine integrative "Wir"-Sprache anstelle einer konfrontativen "Wir gegen die anderen"-Rhetorik zu

verwenden und häufig Mehrdeutigkeiten zu minimieren. Inklusivität, auch in der Sprache, und die Anerkennung von Vielfalt sind wichtige Elemente von Community-Richtlinien und Verhaltensregeln. Solche Praktiken sind besonders wichtig für neurodivergente Nutzer*innen, für die eine explizite Kommunikation Ängste und Fehlinterpretationen reduziert.

Um unerwünschte Infiltration zu vermeiden, können Communities ein grundlegendes Überprüfungsverfahren für neue Mitglieder einführen. Beispielsweise könnte es einen Onboarding-Schritt geben, bei dem neue Mitglieder den Zweck, die Werte und die Grenzen des Raums anerkennen müssen, bevor sie teilnehmen und vollen Zugang erhalten,

Letztendlich ist jedoch das Verhalten der entscheidende Faktor. Die Mitglieder der Community beobachten, wie Moderation in der Praxis funktioniert, und beurteilen die Sicherheit anhand der Schnelligkeit, Konsistenz und Transparenz der Reaktionen auf Verstöße gegen die Community-Regeln, Wenn Hassreden toleriert werden oder die Durchsetzung willkürlich erscheint, verlieren schriftliche Regeln an Glaubwürdigkeit und das Gefühl der Sicherheit schwindet schnell.

Im Fallen von Konflikten innerhalb einer Community sollten Plattformen, anstatt sich ausschließlich und sofort auf Melde- und Sperrungsinstrumente zu verlassen, ihre Mitglieder dazu ermutigen, Konflikte zunächste durch Fairness und konstruktive Gespräche zu lösen und so die Rolle des Einzelnen in der Community zu bekräftigen. Wenn eine Lösung jedoch nicht möglich ist, sollten Disziplinarmaßnahmen ergriffen werden.

➔ UNSERE EMPFOHLENE CHECKLISTE

"MUST-HAVE" / MINDESTANFORDERUNGEN

- ◯ Klare, eindeutige und zugängliche Community-Richtlinien und -Regeln sowie Interventionen bei Verstößen gegen diese
- ◯ Strenge Datenschutzrichtlinien
- ◯ Eingeschränkte Screenshot-Funktion innerhalb privater Communities
- ◯ Zuverlässige Tools zum Melden schädlicher Inhalte
- ◯ Optionale Anonymität
- ◯ Aufsicht für Moderator*innen und Administrator*innen
- ◯ Menschliche Moderator*innen, die Zugang zu psychologischer
- ◯ Unterstützung und Ressourcen haben, wenn sie mit überfordernden Inhalten konfrontiert werden
- ◯ Inklusive Sprache
- ◯ Reaktive und proaktive Maßnahmen gegen schädliche Äußerungen oder Aktivitäten
- ◯ Bei mehrsprachigen Plattformen müssen alle Elemente ordnungsgemäß in alle Sprachen (insbesondere lokale Sprachen) übersetzt sein, einschließlich Nutzungsvereinbarungen, Verhaltensregeln usw.
- ◯ Eigene interne Cybersicherheitskapazitäten, insbesondere für vulnerable Communities

"NICE TO HAVE"

- ◯ Überprüfung während des Onboarding-Prozesses
- ◯ In die Plattforminfrastruktur integrierte Datenschutz- und Schutzmaßnahmen
- ◯ Ressourcen zu Bildungsangeboten rund um Verhalten und Etikette in digitalen Räumen
- ◯ Partizipatives Co-Design, das die Bedürfnisse und Präferenzen der Community berücksichtigt
- ◯ Kund*innendienstmitarbeiter*innen und Moderator*innen, die Mitglieder Community sind oder zumindest über Kenntnisse des kulturellen Kontexts verfügen und die Zeit und emotionale Kapazität haben, um gut durchdachten empathischen Support zu leisten
- ◯ Echtzeit- und direkte menschliche Unterstützung, z.B eine über Zeitzonen hinweg verfügbare Hotline
- ◯ Spezielle algorithmische Tools zur Durchsetzung spezifischer lokaler Normen, auch in Minderheiten- oder lokalen Sprachen
- ◯ Strukturierte digitale Hygiene, z.B. Ruhe- oder Nachtzeiten, in denen Moderator*innen und Administrator*innen über Tools verfügen, mit denen sie beispielsweise die Nachrichtenanzahl auf eine pro Minute beschränken können

RED FLAGS

- ◯ Uneinheitliche Anwendung von Regeln, Zensur vielfältiger, nicht schädlicher Ideen
- ◯ Anhaltende Nichtbeachtung von Verstößen und Schäden, die trotzdem als "freie Meinungsäußerung" zugelassen werden
- ◯ Verpflichtende Offenlegung der tatsächlichen Identität, des Klarnamen
- ◯ Performative Sicherheit, z.B. rein standardmäßiges Kopieren und Einfügen von Community-Regeln
- ◯ Ungeschützte technische Infrastruktur, z.B. Anfälligkeit für Infiltration und Datenlecks
- ◯ Verlass auf Algorithmen oder allgemeine Nutzungsbedingungen zur Bewältigung zwischenmenschlicher Konflikte

EDGELANDS

edgelands.institute