EDGELANDS

CRIANDO ESPAÇOS DIGITAIS SEGUROS COM EMPATIA: UM GUIA PRÁTICO

CONTEXTO

Este documento é resultado de oito semanas de trabalho colaborativo feito por uma equipe de pesquisadores, mentores e artistas, com contribuições de palestrantes convidados, como parte da fase "Pop-Down and Beyond" do Edgelands Institute.

Este guia prático oferece um conjunto de ferramentas e recomendações para o design de espaços digitais seguros, com base em análises de conteúdo de fontes primárias, estudos de caso e pesquisa qualitativa. Esperamos que este guia possa ser utilizado por desenvolvedores, pesquisadores, artistas e pelos próprios usuários dos espaços digitais.

Se você quiser se aprofundar nas complexidades em torno da segurança digital, os atores envolvidos e as definições de espaços digitais seguros e segurança online, você pode acessar a versão estendida deste guia aqui (disponível apenas em inglês).

LANDSLANDS
EDGEEDGE
LANDSLANDS

AGRADECIMENTOS

Este guia foi desenvolvido por Pulkit Mogra, Tatiana Lysova, Lilian Olivia Otero, Catherine Keegan, Nina Martin, Mmabatho Oke, Jessica McClearn e Giovanna da Custódia, sob a orientação de Nina Baranowska, Daniel Odongo, Virgginia Laborão, Vanessa Gathecha e Laura García Vargas.

Gostaríamos de agradecer a todas as pessoas que responderam à nossa pesquisa e contribuíram para a construção das recomendações sobre como criar um espaço digital seguro.

Também agradecemos especialmente às pessoas convidadas que participaram das sessões principais do Research Sprint, compartilhando seus conhecimentos e inspirações.

O design e o layout visual foram criados por Flavia Lozano e Larissa Oliveira.

RESUMO EXECUTIVO

Este guia sintetiza insights feitos a partir de análises de conteúdo, estudos de caso e uma pesquisa qualitativa para fornecer recomendações para o design de espaços digitais seguros que priorizem a empatia e as necessidades específicas da comunidade em vez de abordagens centradas na plataforma.

➔ O PROBLEMA

Ao contrário dos espaços físicos seguros, os ambientes digitais carecem de indicadores de segurança bem definidos. Três atores principais moldam a segurança online: plataformas, governos e comunidades. Mas as abordagens atuais ainda se mostram insuficientes.

AS PLATAFORMAS são projetadas para um "usuário médio" universalizado (tipicamente homens cisgêneros, heterossexuais, brancos e de classe média alta do Norte Global), tornando invisíveis as vulnerabilidades de populações marginalizadas. A lógica corporativa prioriza as métricas de engajamento em detrimento do bem-estar do usuário.

AS REGULAMENTAÇÕES GOVERNAMENTAIS focam estritamente na prevenção de "danos tangíveis" (material de abuso infantil, terrorismo, fraude), ignorando a natureza subjetiva e contextual da sensação de segurança online.

AS COMUNIDADES exercem um papel crucial na governança da segurança por meio de códigos de conduta e moderação voluntária. Porém, isso é frágil na maioria das vezes, pois permanecem estruturalmente subordinadas às arquiteturas das plataformas.

➔ PRINCIPAIS RESULTADOS DE PESQUISA

Nossa análise identificou quatro condições fundamentais para a segurança digital:

CONDIÇÕES RELACIONAIS: respeito aos limites pessoais sem a necessidade de conflitos, moderação eficaz como um trabalho relacional (e não apenas como aplicação de regras) e redução do custo emocional vindo de situações em que os usuários precisam se defender constantemente.

CONDIÇÕES CULTURAIS: normas sociais de dignidade e não julgamento, inclusão linguística para idiomas e comunidades e redes de apoio que oferecem infraestrutura de proteção para grupos vulneráveis.

CONDIÇÕES PROCESSUAIS: regras claras e aplicadas de forma consistente, autonomia do usuário sobre a visibilidade e o rastreamento de dados e estruturas de governança legítimas baseadas em experiências vividas, em vez de decisões corporativas arbitrárias.

CONDIÇÕES DE INFRAESTRUTURA: confiabilidade e integridade técnica das plataformas, incluindo transparência no armazenamento de dados, mecanismos de reporte eficazes e responsabilização em caso de falhas nos sistemas.

Esses temas são seguidos por recomendações essenciais em níveis — "indispensáveis", "desejáveis" e "sinais de alerta" — para a criação de espaços digitais mais seguros.

➔ CONCLUSÃO

Este guia serve tanto como um recurso prático quanto como um apelo para reimaginar os espaços digitais como ambientes governados pela comunidade, onde a segurança é co-criada por aqueles que habitam nesses espaços — e não imposta por arquiteturas corporativas otimizadas para o engajamento em detrimento do bem-estar. Reconhecemos as limitações de nossa pesquisa, incluindo a necessidade de perspectivas mais técnicas, maior envolvimento da comunidade e tradução para línguas indígenas. Por fim, defendemos mais pesquisas aprofundadas sobre as questões relacionadas ao tema da segurança online.

COMO IDENTIFICAR, CRIAR E GERENCIAR ESPAÇOS DIGITAIS SEGUROS: UM GUIA PRÁTICO

Confira abaixo um guia prático para espaços digitais seguros, baseado em nossa pesquisa.

→ REQUISITOS OBRIGATÓRIOS

POLÍTICA E DOCUMENTAÇÃO

Uma documentação clara e bem elaborada desempenha um papel crucial na sensação de segurança em espaços digitais. A segurança aumenta quando as regras são claras, explícitas, acessíveis e visíveis antes de ingressar em um grupo. Essa transparência permite que os indivíduos tomem decisões informadas sobre se um espaço digital está alinhado com seus valores e necessidades e, consequentemente, se desejam ou não participar dele.

As regras devem articular claramente os comportamentos aceitáveis, as normas de comunicação e as respectivas consequências. Isso contribui para a sensação de segurança, reduzindo a incerteza e reforçando a compreensão de que comportamentos prejudiciais serão desencorajados. Além disso, essa articulação deve ser feita em linguagem acessível a todos os membros da comunidade.

Boas práticas em regras comunitárias abordam e proíbem explicitamente assédio, bullying, discurso de ódio, conteúdo extremista e formas de abuso de identidade (por exemplo, falsificação de identidade, uso de IA para criar imagens impróprias de alguém sem o seu consentimento, etc). Se as regras forem um tanto ambíguas, deve haver espaço para discussão, revisão e reformulação. Substituir linguagem ampla ou abstrata, como comunicação "ecológica" ou

"positiva", por expectativas explícitas sobre respeito, não julgamento e transparência ajuda a tornar as normas mais compreensíveis e aplicáveis.

As diretrizes comunitárias devem ser elaboradas com a participação das próprias comunidades. Além disso, devem ser revisadas de forma contínua e iterativa, levando em consideração as experiências dos membros da comunidade no terreno.

A governança de dados deve ser claramente definida, esclarecendo como os dados pessoais são coletados, usados, armazenados e protegidos. Práticas éticas de dados, princípios de privacidade desde a concepção até a comunicação transparente e acessível sobre as medidas de proteção de dados são essenciais para a sensação de segurança em ambientes digitais. Se uma ferramenta digital for multilíngue, é fundamental que todos os elementos, incluindo aqueles relacionados à proteção de dados, sejam traduzidos para todos os idiomas.

Por fim, as regras só são eficazes quando aplicadas de forma consistente e igualitária a todos, incluindo moderadores e administradores. Quando a aplicação das regras é desigual ou percebida como arbitrária, a confiança se deteriora rapidamente e os sentimentos de exclusão ou vulnerabilidade aumentam.

REQUISITOS TÉCNICOS

Ferramentas de segurança robustas e atualizadas são componentes vitais da segurança digital, mas, simultaneamente, devem respeitar a privacidade, ser sensíveis ao contexto e não punitivas. Elementos técnicos essenciais que contribuem para a sensação de segurança incluem filtros de conteúdo, remoção automática de

conteúdo perturbador ou ilegal e mecanismos de denúncia. Mecanismos de criptografia e anonimização, prevenção de vazamento de dados e medidas de segurança atualizadas, como autenticação de dois fatores ou proteção contra vazamento de dados, também são cruciais para gerar a sensação de segurança em espaços digitais.

No entanto, a moderação automatizada muitas vezes extrapola seus limites, removendo conteúdo inofensivo devido à falta de contextualização ou a regras excessivamente rígidas. Tal extrapolamento pode limitar a participação, silenciar expressões legítimas e minar a confiança na plataforma. Portanto, a moderação automatizada deve funcionar em conjunto com a supervisão humana. Além disso, deve ser possível entrar em contato com a plataforma e contestar uma moderação automatizada excessivamente rigorosa, fornecendo contextualização e explicações para o conteúdo.

As proteções técnicas também devem abranger a prevenção de abusos por parte da comunidade, o uso indevido de dados privados e a presença de usuários hostis em comunidades sensíveis ou vulneráveis. Os recursos de segurança são mais eficazes quando limitam esses riscos de forma proativa, em vez de serem aplicados de forma reativa após a ocorrência de danos.

Os usuários devem ter controle sobre sua própria privacidade em uma plataforma. As plataformas devem fornecer ferramentas para gerenciar a visibilidade de seus perfis, divulgar informações pessoais seletivamente e escolher entre modos de participação públicos e privados. Esses recursos concedem aos usuários um maior nível de controle sobre sua exposição, de acordo com suas necessidades e riscos.

Políticas de uso de nome real podem ser prejudiciais quando o anonimato e a privacidade são importantes para a segurança. Para muitos usuários, especialmente aqueles de comunidades marginalizadas ou em contextos politicamente sensíveis, o anonimato não é uma preferência, mas uma medida de proteção vital.

Por fim, as plataformas devem incorporar de forma ativa e preventiva ajustes e opções para indivíduos ou comunidades com necessidades específicas, tornando a acessibilidade como norma, em vez de um luxo ou algo pelo qual seja preciso argumentar.

ELEMENTOS DA COMUNIDADE

Na ausência de limites físicos, os usuários se baseiam em um conjunto de pistas digitais informais, porém poderosas, para avaliar se um espaço é seguro. Eles podem procurar marcadores visuais — efetivamente lendo a "linguagem corporal" da plataforma; pronomes em perfis, símbolos de inclusão, avisos de conteúdo ou Códigos de Conduta fixados no topo do feed funcionam como sinais imediatos de que existem limites e que o espaço está sendo cultivado intencionalmente.

Esses sinais visuais são reforçados por pistas linguísticas que moldam a atmosfera interpessoal de uma comunidade. Espaços seguros tendem a adotar uma linguagem inclusiva com o pronome "nós", em vez de uma retórica conflituosa do tipo "nós contra eles", e frequentemente utilizam indicadores de tom para minimizar ambiguidades. A inclusão por meio da linguagem e o reconhecimento da diversidade são elementos importantes para as diretrizes, as regras e os códigos de conduta da comunidade. Essas

práticas são particularmente importantes para usuários neurodivergentes, para os quais a comunicação explícita reduz a ansiedade e as interpretações errôneas.

Para evitar infiltrações indesejadas, as comunidades podem introduzir um processo básico de verificação para novos membros. Por exemplo, pode haver uma etapa de integração exigindo que os novos membros reconheçam o propósito, os valores e os limites do espaço antes de participar e ganhar acesso total.

Em última análise, porém, o indicador mais decisivo é o comportamental. Os membros da comunidade observam como a liderança e a moderação operam na prática, avaliando a segurança pela rapidez, consistência e transparência das respostas a violações das regras da comunidade. Quando o discurso de ódio é tolerado ou a aplicação das regras parece arbitrária, as regras escritas perdem credibilidade e a sensação de segurança se dissipa rapidamente.

Em caso de conflito dentro de uma comunidade, em vez de depender exclusivamente e imediatamente de ferramentas de denúncia e banimento, as plataformas devem incentivar os membros a resolverem os conflitos primeiramente por meio da gentileza e do diálogo construtivo, reafirmando assim o papel do indivíduo na comunidade. No entanto, medidas disciplinares devem ser aplicadas rapidamente quando a resolução não for possível.

➔ NOSSA LISTA DE VERIFICAÇÃO RECOMENDADA

REQUISITOS OBRIGATÓRIOS

- ◯ Diretrizes e regras claras, explícitas e acessíveis, com intervenções previstas caso elas sejam infringidas
- ◯ Política rigorosa de proteção de dados
- ◯ Funcionalidade de captura de tela restrita em comunidades privadas
- ◯ Ferramentas confiáveis para denunciar conteúdo perturbador
- ◯ Opções de anonimato
- ◯ Supervisão de moderadores e administradores
- ◯ Moderadores humanos, com acesso a suporte psicológico caso se deparem com conteúdo perturbador
- ◯ Linguagem inclusiva
- ◯ Medidas proativas contra discursos de ódio ou atividades prejudiciais
- ◯ Para plataformas multilíngues, todos os elementos devem estar devidamente traduzidos para todos os idiomas (especialmente os locais), incluindo termos de uso, códigos de conduta, etc.
- ◯ Capacidade interna dedicada à cibersegurança, especialmente para comunidades em risco.

BONS REQUISITOS ADICIONAIS

- ◯ Verificação do perfil antes de entrar na comunidade
- ◯ Medidas de privacidade e proteção integradas à infraestrutura da plataforma
- ◯ Recursos educacionais sobre comportamento e etiqueta em espaços digitais
- ◯ Design participativo que incorpora as necessidades e as preferências da comunidade.
- ◯ Agentes de suporte/moderadores que sejam membros da comunidade de usuários (ou que tenham conhecimento do contexto cultural) e que disponham de tempo e capacidade emocional para fornecer suporte ponderado
- ◯ Suporte humano direto e em tempo real (por exemplo, uma linha de ajuda disponível em diferentes fusos horários).
- ◯ Ferramentas algorítmicas específicas para aplicar normas locais, inclusive em línguas minoritárias ou locais.
- ◯ Higiene digital estruturada, por exemplo, horários de menor movimento ou noturnos, nos quais moderadores e administradores podem utilizar ferramentas como limitar o envio de mensagens a uma por minuto.

SINAIS DE ALERTA

- ◯ Inconsistência na aplicação das regras, censura de diversas ideias inofensivas.
- ◯ Falta persistente de resposta a violências ou discurso de ódio permitido como "liberdade de expressão".
- ◯ Exposição obrigatória da identidade real
- ◯ Segurança "performativa", como apenas copiar e colar regras da comunidade por padrão.
- ◯ Infraestrutura técnica desprotegida, como exposição a infiltrações e vazamentos de dados.
- ◯ Recorrer a algoritmos ou termos de serviço genéricos para lidar com conflitos interpessoais.

EDGELANDS

edgelands.institute