Beyond Filtered Reality

Case Study

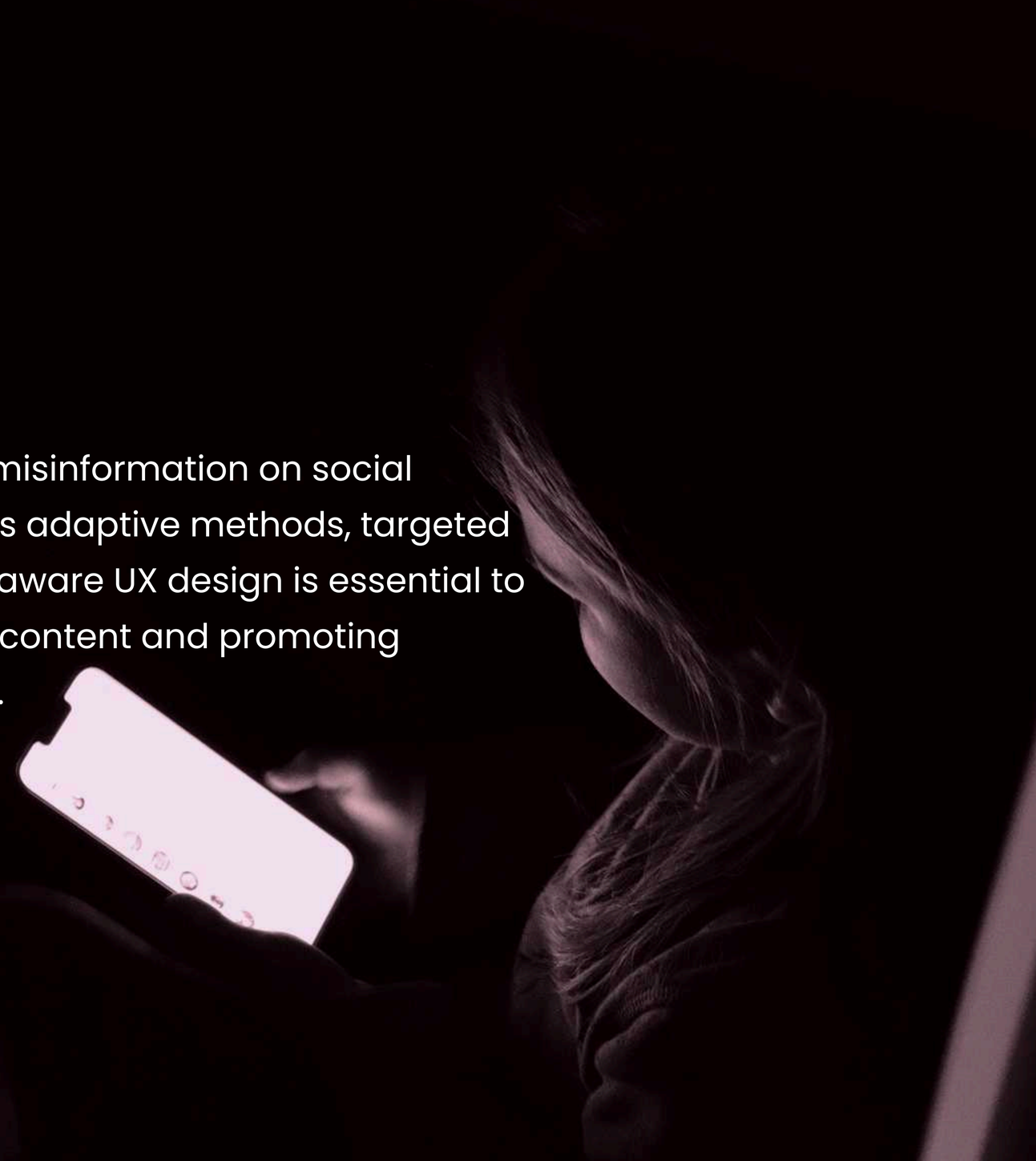
Designing User centric interfaces to Tackle the challenges of AI - Propagated Misinformation on Instagram

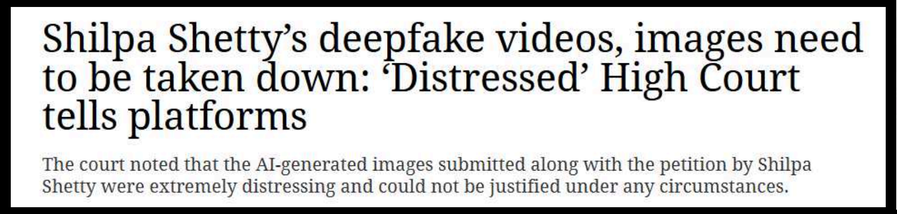
Brief

In the age of artificial intelligence, misinformation is becoming increasingly difficult to identify and increasingly easy to spread. Generative AI tools and deepfakes now blur the boundary between what is real and what is fabricated, allowing false content to circulate at unprecedented speed and scale. As social media platforms such as Twitter, Instagram, Facebook, WhatsApp, and Telegram continue to function as primary sources of information, the responsibility to detect, limit, and prevent the spread of misinformation has become more urgent than ever.

The Gap

A significant gap persists in addressing AI-generated misinformation on social platforms. Current UX solutions fail to keep pace with its adaptive methods, targeted personalization, and echo chamber reinforcement. AI-aware UX design is essential to equip users with tools for critically assessing synthetic content and promoting responsible engagement in this dynamic environment.




Shilpa Shetty's deepfake videos, images need to be taken down: 'Distressed' High Court tells platforms

The court noted that the AI-generated images submitted along with the petition by Shilpa Shetty were extremely distressing and could not be justified under any circumstances.


Press release

Online disinformation : UNESCO unveils action plan to regulate social media platforms

Audrey Azoulay, Director-General of UNESCO sounded the alarm on Monday about the intensification of disinformation and hate speech online, which constitutes "a major threat to stability and social cohesion". To put an end to this scourge she unveiled UNESCO's action plan, the result of extensive worldwide consultations and is backed by a global opinion survey underlining the urgent need for action.


THE PULSE | SOCIETY | SOUTH ASIA

India's Growing Misinformation Crisis: A Threat to Democracy

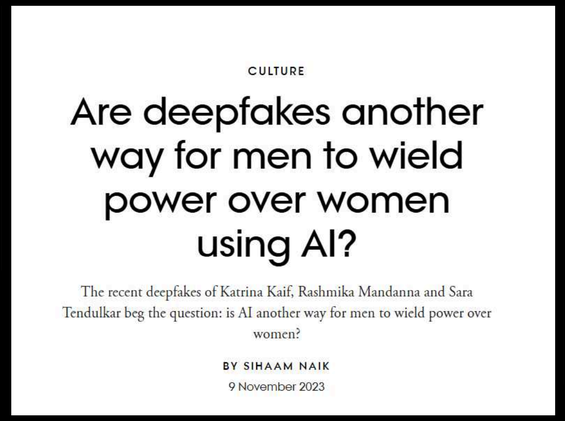
Failure to curb dissemination of falsehood via social media can have long-term effects on democratic processes.

By Abhinav Mehrotra and Amit Upadhyay
March 19, 2025


NEWS > LOCAL NEWS > I-TEAM INVESTIGATIONS

AI deepfakes fuel surge in online scams targeting families

Cybersecurity expert warns deepfake technology will become primary scam tool by late 2026


CULTURE

Are deepfakes another way for men to wield power over women using AI?

The recent deepfakes of Katrina Kaif, Rashmika Mandanna and Sara Tendulkar beg the question: is AI another way for men to wield power over women?

BY SIHAAM NAIK
9 November 2023


Poll Shows Most U.S. Adults Think AI Will Add to Election Misinformation in 2024

Online Disinformation

UNESCO unveils action plans to regulate social media platforms

War Borne

War propaganda can be turbocharged from either sides

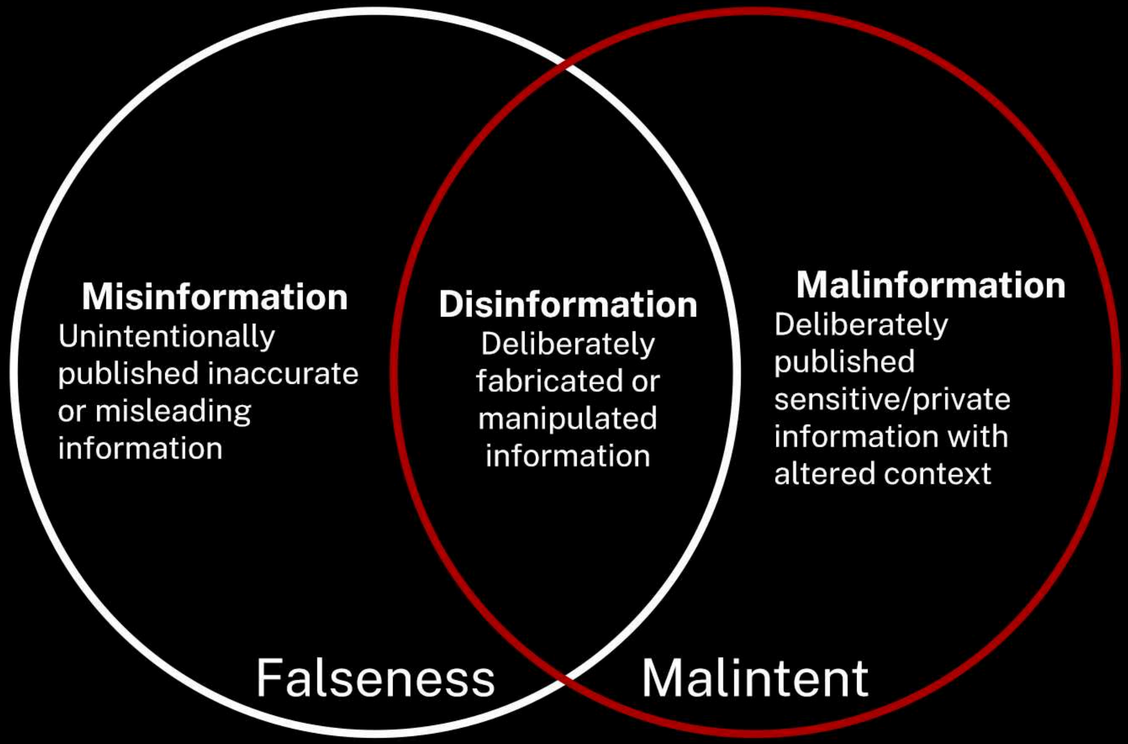
Women safety

Deepfakes can be detrimental to the safety of women

Upcoming election

Remember Cambridge analytics? well.....

Information Disorder



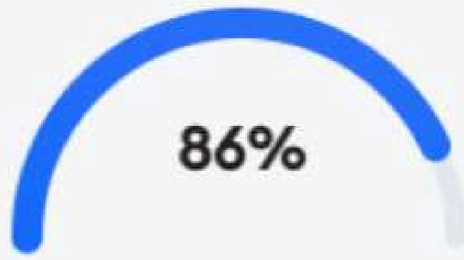
To understand and study the complexity of the information ecosystem, we need a common language. The current reliance on simplistic terms such as 'fake news' hides important distinctions and denigrates journalism. It also focuses too much on "true" vs "fake", whereas information disorder comes in many shades of "misleading".

AI Democratization

AI democratization describes the movement to make AI technologies, tools, and capabilities widely accessible to diverse users. It focuses on reducing technical barriers, empowering non-experts—individuals and organizations alike—to harness AI effectively for their unique needs and applications.

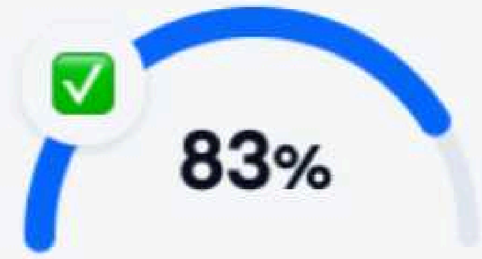
Mainstream Technology

52% of businesses accelerated AI adoption,
and 86% claimed it was already a "mainstream technology."



Source: Gartner

83% of companies consider using
AI to be a high priority



Source: Forbes

of consumers are concerned about misinformation from AI.
Forbes

This is leading to the democratization of disinformation & misinformation

The democratization of AI, which refers to the increasing accessibility and usability of AI technologies, has also led to the democratization of misinformation and disinformation. This is because AI-powered tools can be used to create and spread false or misleading information more easily and effectively than ever before.

Threats

01 Disrupt Democracy & Society

02 Erode Trust in Institutions

03 Invades Individual Privacy & Autonomy

04 Lead to violence & unrest

05 Undermine Public Health

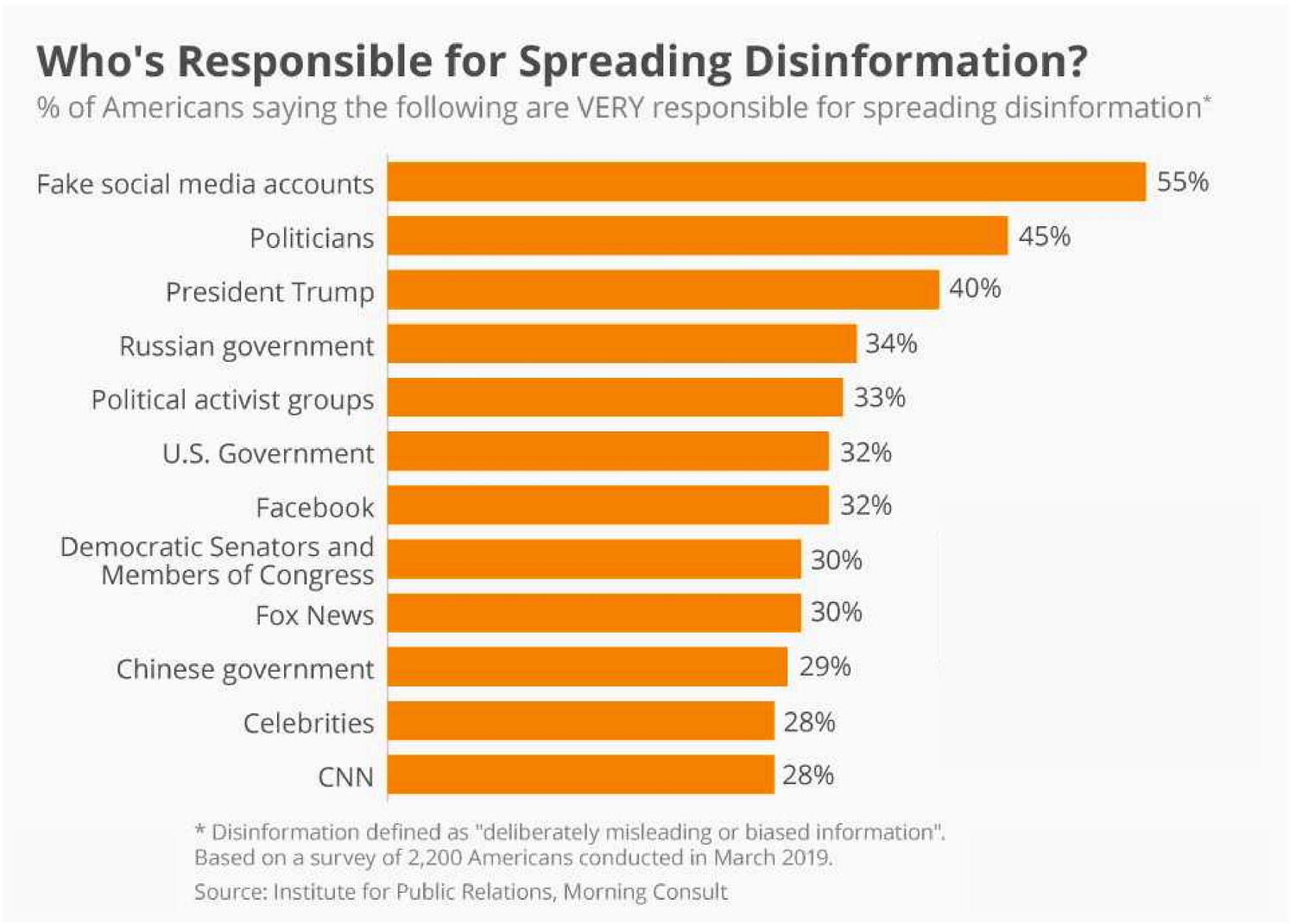
06 Challenges Critical Thinking

07 Social Polarization

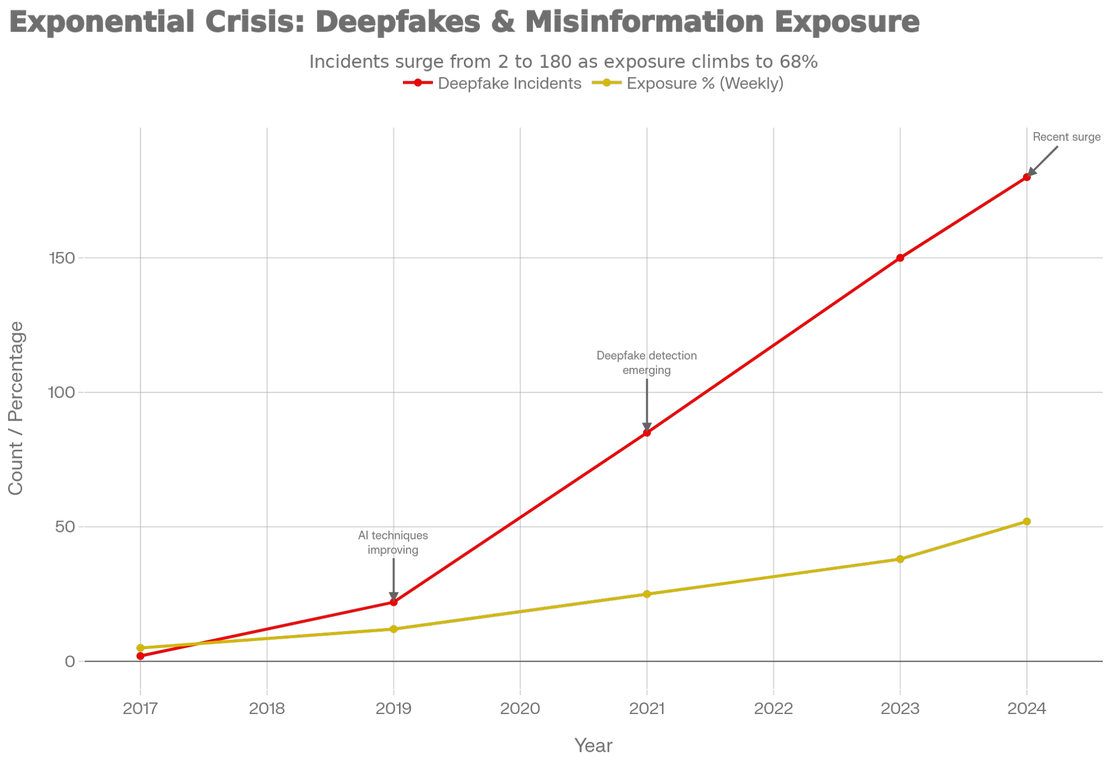
08 Incite Extremism

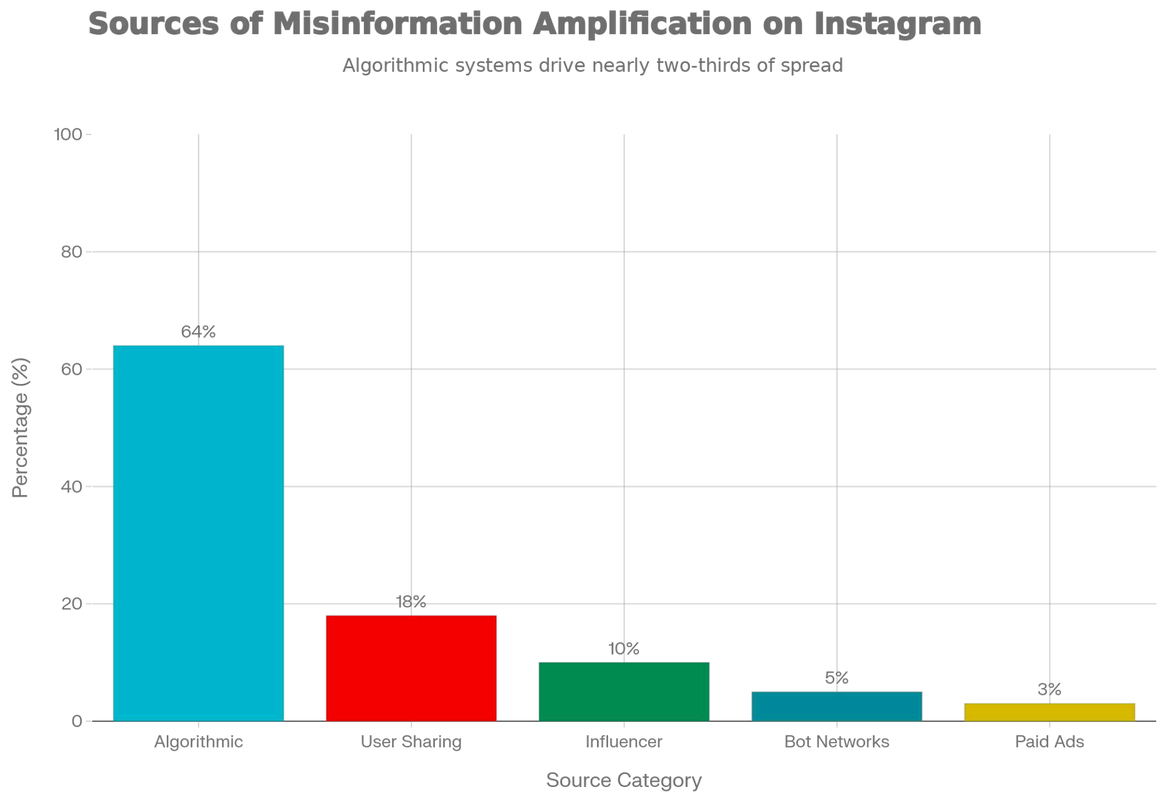
09 Normalizing Deception & Manipulation

Impacts on Social Media



"While AI has undoubtedly played a role in the dissemination of misinformation, it is important to recognize that it is not the sole factor."





Why is this a UX Problem??

This issue is a UX problem because the solution lies in UX principles and interface design. Addressing misinformation on social media requires a UX-oriented approach that considers the user experience and the factors that make users vulnerable to misinformation.

Designing for Trustworthiness

Social media platforms should prioritize transparency, accountability, and fact-checking mechanisms to build user trust and confidence.

User-Centric Content Moderation

Content moderation strategies should consider the user experience, avoiding overly restrictive measures that stifle legitimate expression.

Mindful Engagement

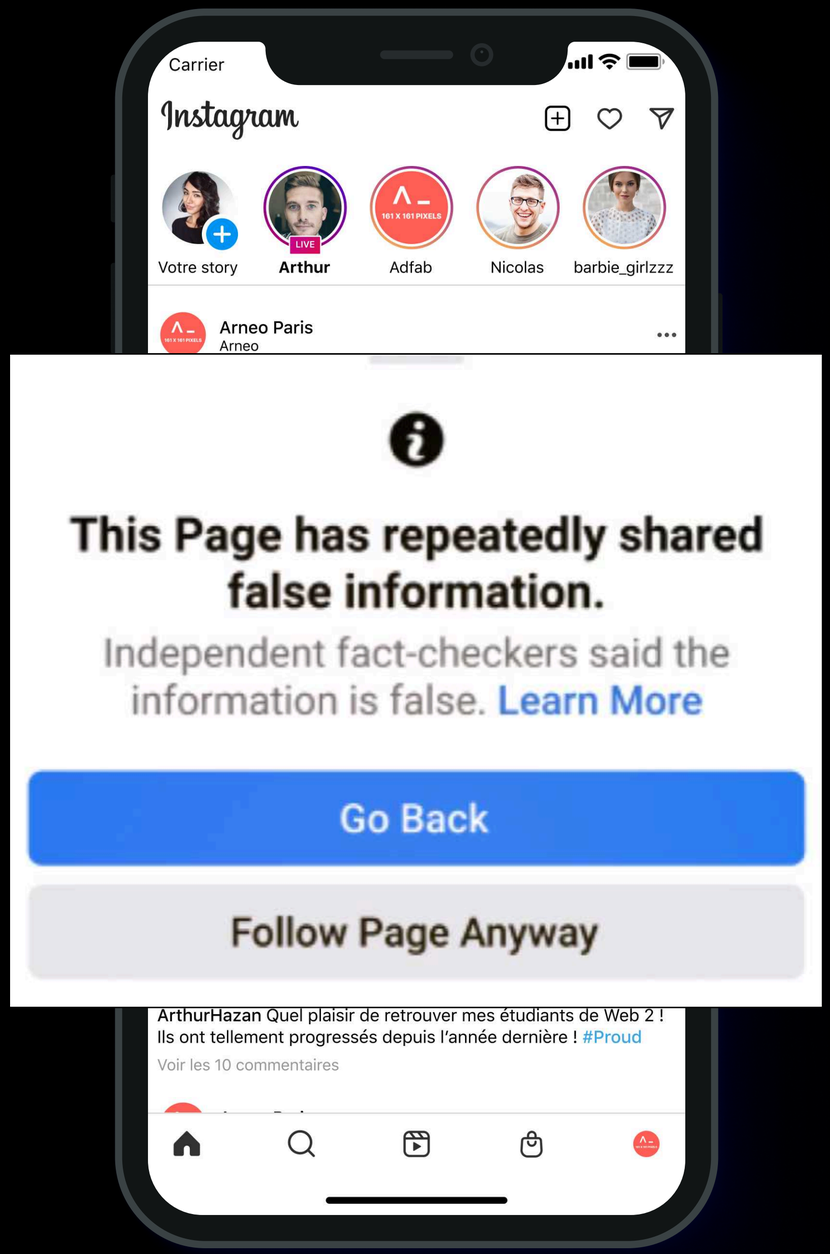
From designs ensuring more user engagement with the content, to designing for mindful engagement, an UX paradigm shift plays a huge role in this issue.

Current UX Interventions

Information-prompts and warning

Fact-checking and verification tools

Content moderation and removal



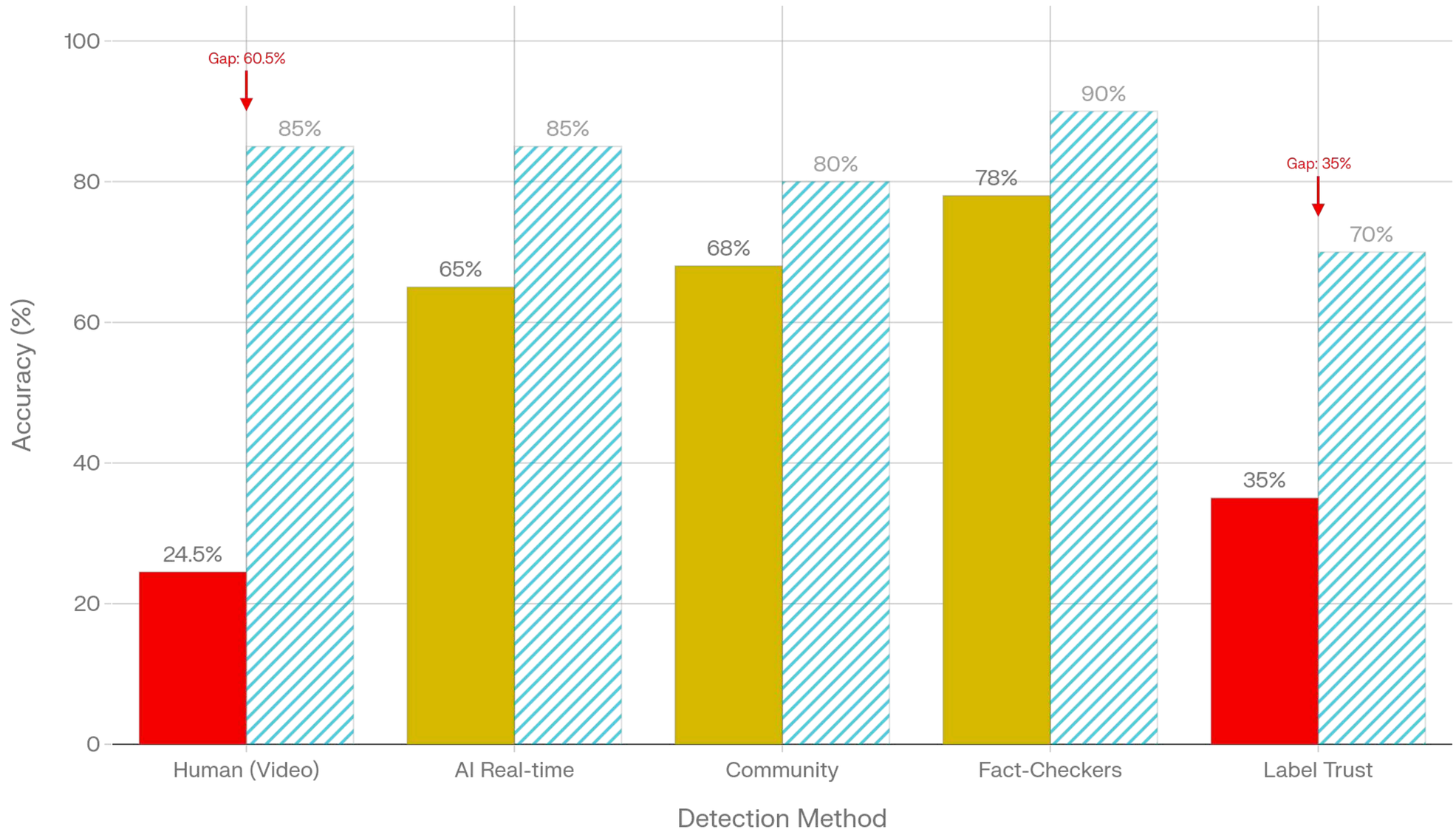
Platform wise UX Interventions

Platform	UX Interventions	Strengths	Limitations
Facebook	Fact-checking labels, third-party fact-checker partnerships, reduced distribution of misleading content, warnings before sharing potentially misleading posts	Wide reach of fact-checking labels, collaboration with reputable fact-checkers	Limited ability to identify and address all forms of misinformation, potential for over-blocking of legitimate content
Instagram	Fact-checking labels, third-party fact-checker partnerships, reduced distribution of misleading content, "Fact Check" tag on disputed posts.	Visually prominent fact-checking tags, integration with fact-checking organisations	Similar to Facebook, may not capture all forms of misinformation, potential for over-blocking of genuine content
Twitter	Pop-up prompts for disputed content, warnings before retweeting disputed content, "Misleading" label for disputed tweets, "Get the facts" feature for disputed tweets	Direct prompts and warnings for potentially misleading content, context-specific information about disputed claims	Potential for user fatigue with repeated warnings, may not effectively address all types of misinformation
WhatsApp	Forwarded message limits, chat forwarding labels, fact-checking chatbot, group administrator controls	Reduced spread of misinformation through message forwarding limits, increased transparency through forwarding labels	Limited reach of fact-checking chatbot, reliance on group administrators to curb misinformation

Platform wise Measures of UX Mitigation

Mitigation Parameter	Fact-checking labels	Third-party fact-checker partnerships	Content reduction for misleading posts	User warnings for misleading content	Integration with fact-checking orgs	Transparency about fact-checking efforts	User empowerment
Facebook				Limited		Moderate	Limited
Instagram				Limited		Moderate	Limited
Twitter				More Prominent		Moderate	Moderate
WhatsApp	Limited Implementation		Not Applicable	Limited	Limited	Limited	Moderate

Gaps and Limitations



01
Limited Effectiveness in Addressing All Forms of Misinformation

02
Over-Reliance on Fact-Checking Labels and Warnings

03
Inadequate User Empowerment and Education

04
Lack of Granular User Controls

05
Insufficient Cross-Platform Collaboration

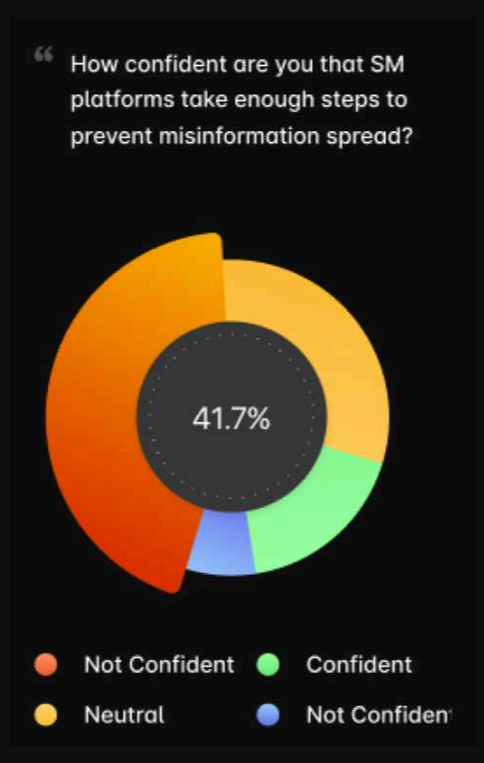
06
Potential for Unintended Consequences

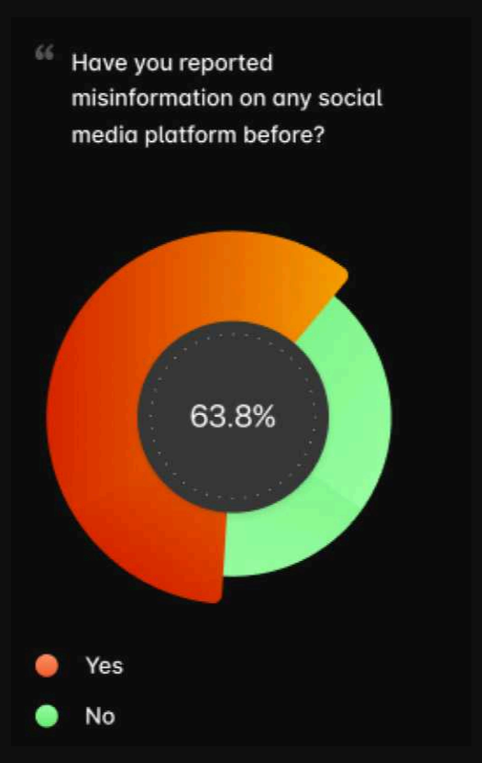
07
Limited User Engagement and Feedback

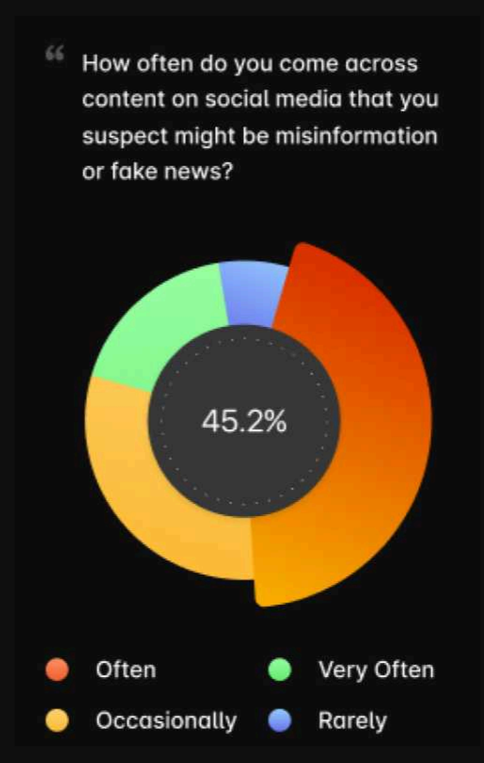
08
Lack of Transparency and Accountability

Quantitative Research

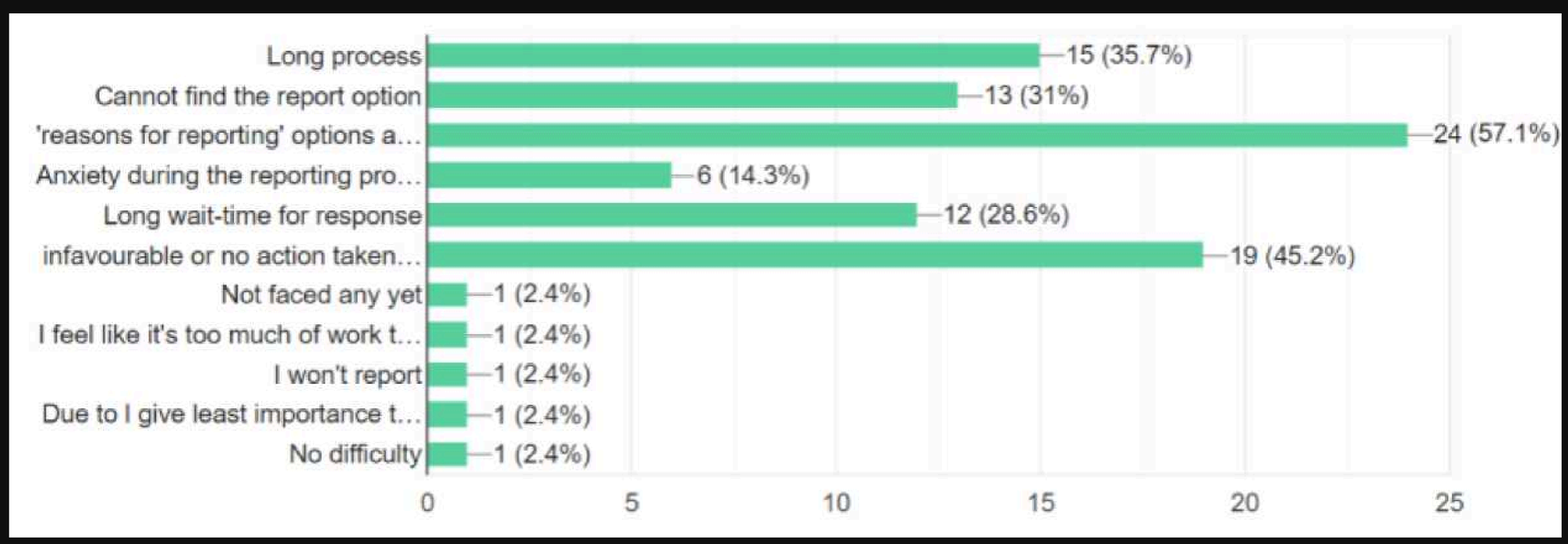
This quantitative research study aims to gather insights from a sample of 50+ people from random backgrounds to understand their perceptions, attitudes, and behaviors towards misinformation on social media.



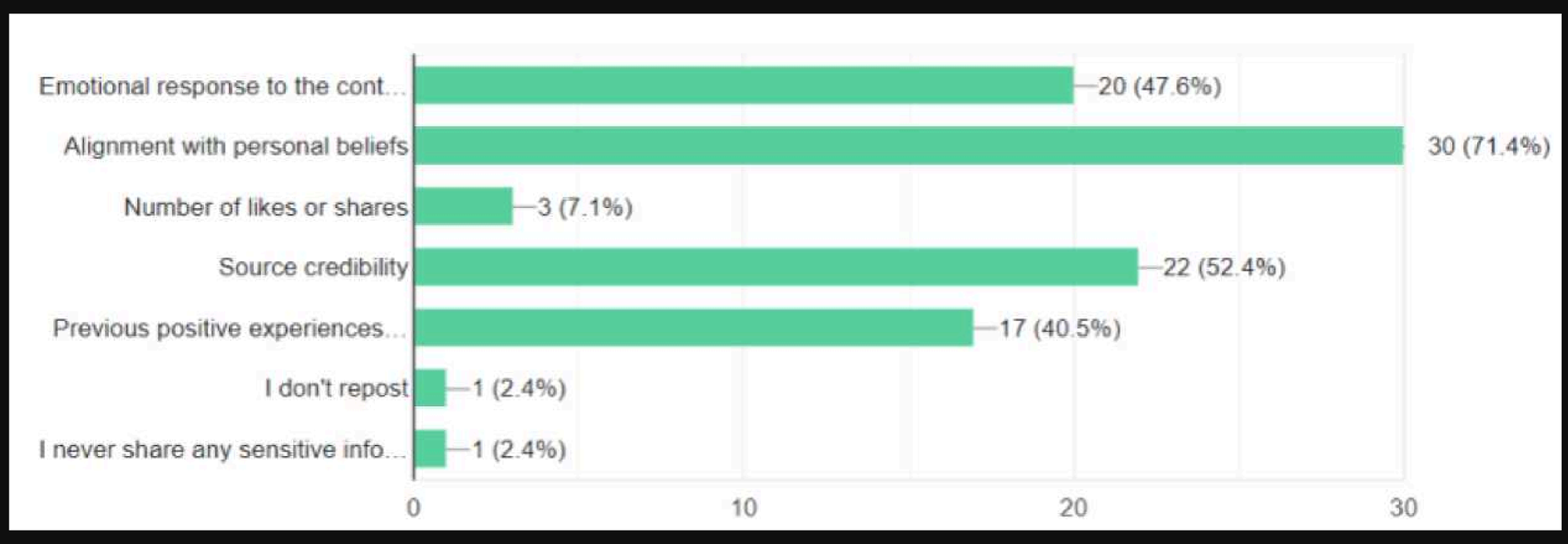




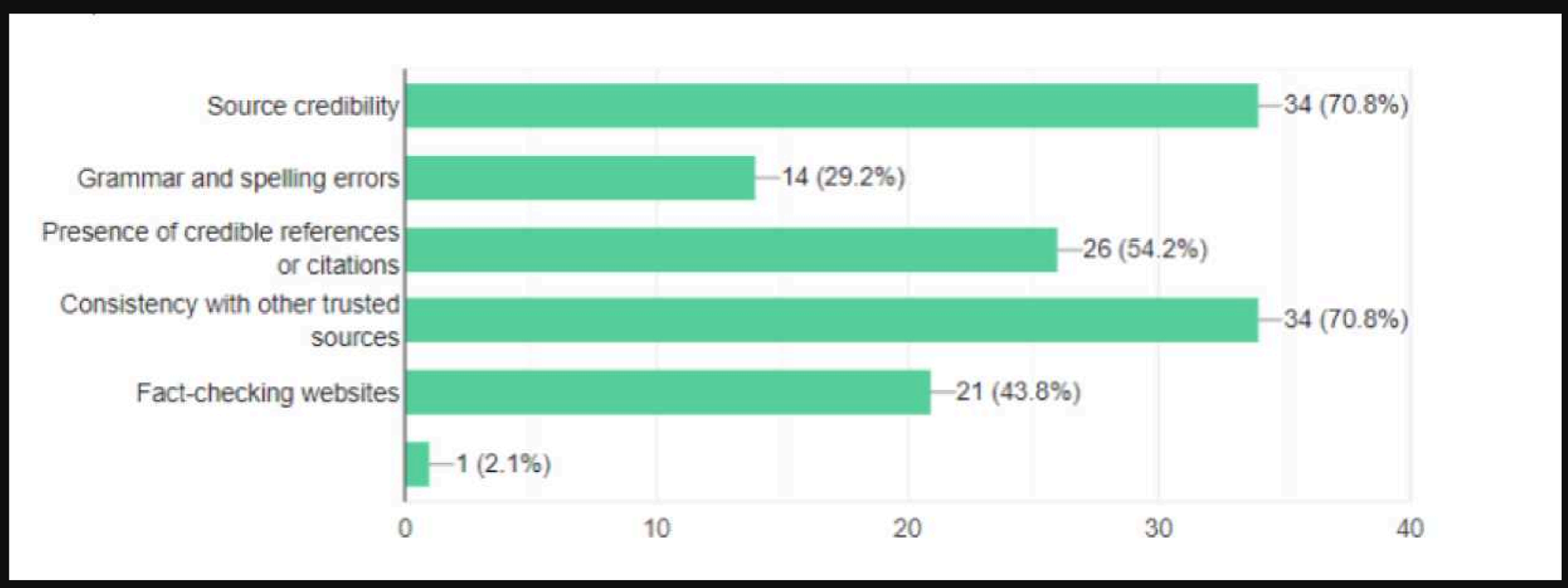
What factors influence your decision to share or repost information on social media?



What difficulties did you face while reporting misinformation?



What cues or indicators do you use to identify misinformation on social media?



Results and Findings

Prevalence and Impact

The survey revealed that a significant majority of respondents (97.6%) have encountered misinformation on social media platforms.

User Challenges

Only 41.4% of respondents expressed confidence in their ability to identify misinformation.

User Perception

A majority of respondents recognized fact-checking labels (68%) and warnings for potentially misleading content (62%), indicating that these interventions have gained some visibility among users. However, their opinions on the effectiveness of these interventions were mixed.

Key Insights

Community Driven Solutions

A significant majority (63.8%) relied on user comments and discussions for discovering fake news.

Report Experience

A significant 53.2% of the sample have stated that they experience anxiety during the reporting process.

Source Indicators

72.3% relied on the source credibility, hence it is important to highlight the sources.

Community Driven Solutions

A significant 40.4% are not very confident while a 36.2% of the respondents are neutral when asked if they about their trust on social media.

Transparency & Accountability

Also 48.9% have said that there is almost no action or unfavorable action taken after reporting

Confirmation Bias

A 70.2% majority have stated that they are likely to repost or share if the post aligns with their personal beliefs.

Implications for UX Design

The findings of this study underscore the importance of UX design in combating misinformation and promoting informed online behaviour. Key UX design principles for addressing misinformation include:

Source Credibility Indicators

By making source credibility more transparent, UX design can help users make informed judgments about the information they encounter.

Address Confirmation Bias

Algorithmically diversifying content feeds, highlighting opposing viewpoints, and providing fact-checking prompts for content that aligns with users' beliefs.

Community-Driven Tools

This could involve implementing user-generated reporting mechanisms, crowd-sourced fact-checking tools, and community-moderated forums.

Promoting Fact-Checking Habits

Introducing delays or prompts before sharing, integrating fact-checking tools into the sharing process, and highlighting counter arguments or dissenting opinions.

Enhance Reporting Mechanisms

Streamlining the reporting process, offering a comprehensive list of reporting options, and providing status updates on reported content.

Root Cause Analysis (5 Whys)

Why # 1

Users Share Unverified Content

Users publish misinformation on Instagram without checking accuracy first.

Why #2

Verification Requires Too Much Friction

Fact-checking demands exiting Instagram, opening a search engine, evaluating sources, and returning to the app. This multi-step process takes 2–5 minutes—making it impractical during typical browsing.

Why #3

No Integrated Verification During Sharing

Fact-checking tools exist on Instagram (community notes, disputed badges) but only appear *after* content is published and flagged. Users don't see credibility signals *during* the sharing decision moment.

Why #4

System Optimizes for Speed, Not Accuracy

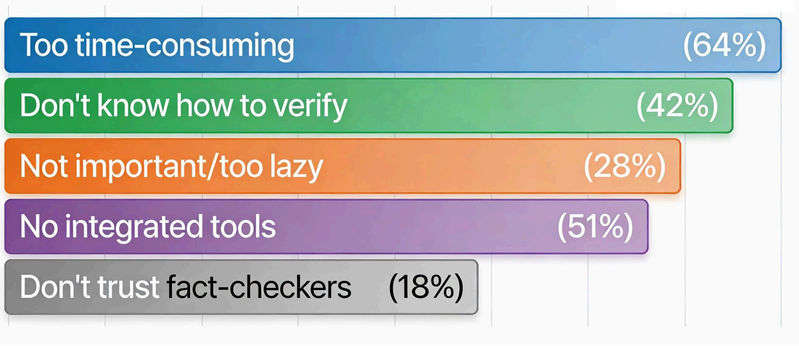
Instagram's architecture prioritizes rapid content propagation to maximize engagement and time-on-platform. Community validation is designed as a moderation tool (removing false content post-hoc) rather than a user empowerment tool (helping verify before sharing).

Why #5

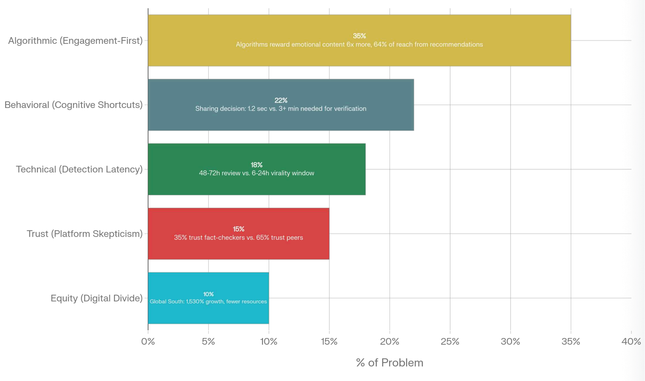
Fundamental Mismatch: Motivation vs. Ability

Research reveals 87% of users want to share responsibly (high motivation), but only 1.5/10 report ability to verify quickly (very low ability). The interaction design lacks community-driven content validation mechanisms that surface credibility signals in-platform before publishing. Users face a binary choice: share without verification or abandon the share due to friction.

Why Users Skip Verification



Root Cause Analysis: Why Misinformation Spreads

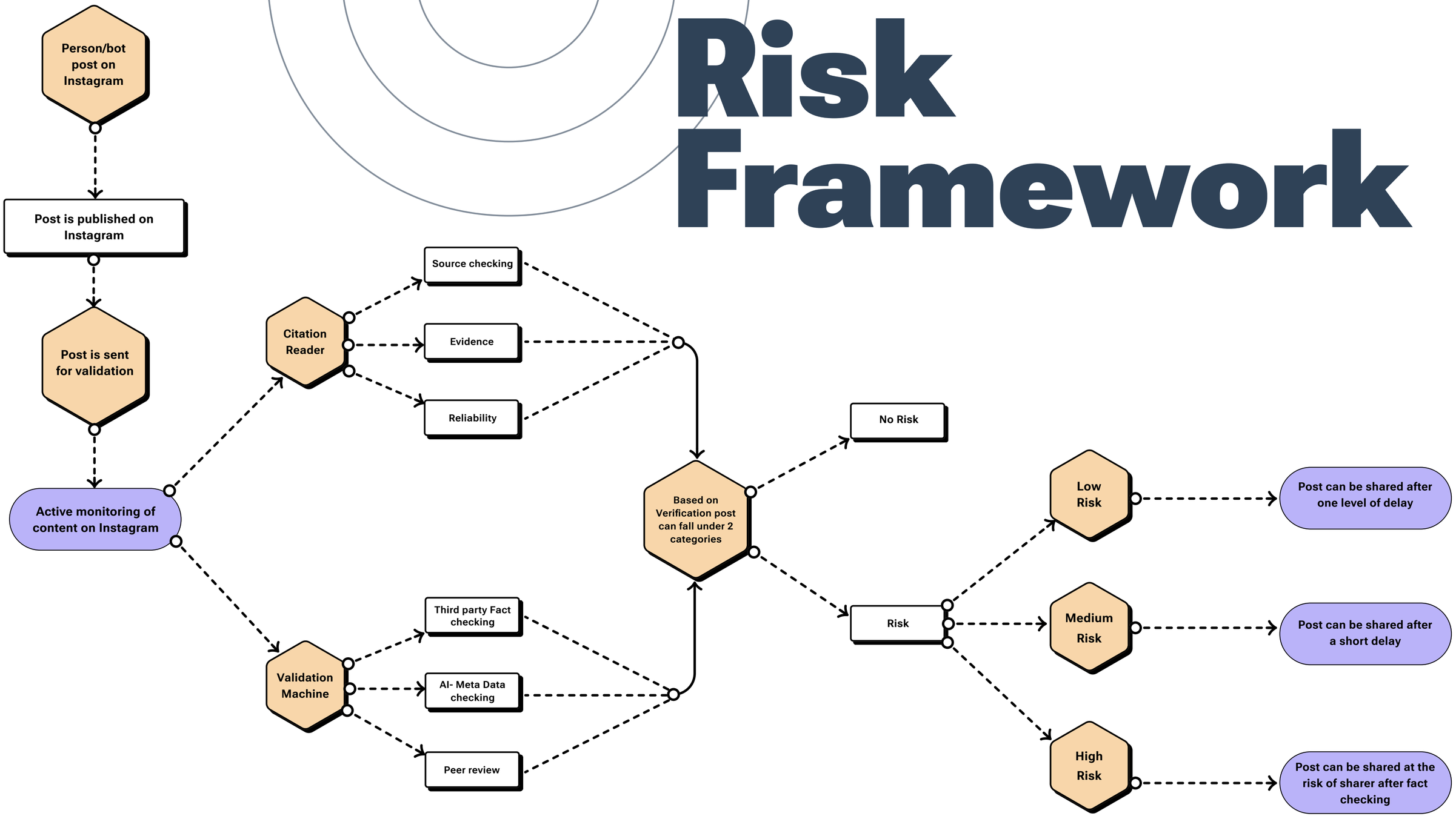


Root cause statement

The core problem is not that users lack motivation to share responsibly—it is that they lack the ability to verify content efficiently within the platform's sharing workflow. Current design relegates community validation to post-publication moderation, not pre-sharing assistance.

Critical Insight: This is fundamentally an ability and design problem, not a motivation or awareness problem. Users don't need more warnings or guilt—they need frictionless, in-platform tools to assess credibility at the moment of sharing

Risk Framework



Problem Statement

How Might We Enable Users to Make Informed Sharing Decisions?

Instagram users want to share credible content but lack integrated tools to assess source accuracy and community consensus before publishing. The current interaction design treats fact-checking as post-hoc moderation rather than pre-sharing assistance, placing the burden of verification on individual users outside the platform. This gap between user motivation (high) and user ability (low) results in widespread unverified sharing.

Redesign Opportunity: Integrate community-driven credibility signals into the sharing workflow, enabling users to validate sources and assess content accuracy in-platform, in real-time, with minimal friction.

User Persona

Amandeep

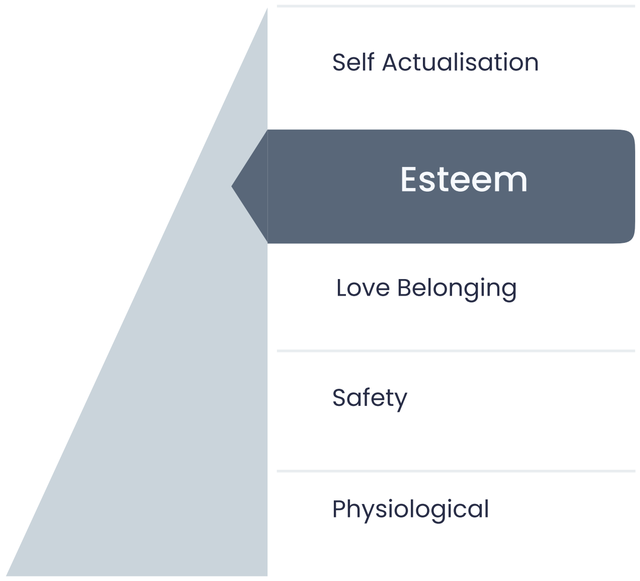
Graphic Designer

About

23 Delhi

B.sc Employee

Maslow Pyramid



Description

Amandeep is a young, tech-savvy individual who is constantly connected to social media. He uses social media to stay up-to-date on current events, connect with friends and family, and discover new trends. Alex is aware of the issue of misinformation on social media, but he Bachelors in Sociology

@ Goals

- To be able to easily identify and avoid misinformation on social media.
- To have access to reliable and trustworthy sources of information.
- To be able to share information confidently without fear of spreading misinformation. feels that he is generally able to identify and avoid it

Frustrations

- To be able to easily identify and avoid misinformation on social media.
- To have access to reliable and trustworthy sources of information.
- To be able to share information confidently without fear of spreading misinformation.

Sunita Jain

Retired Nurse

About

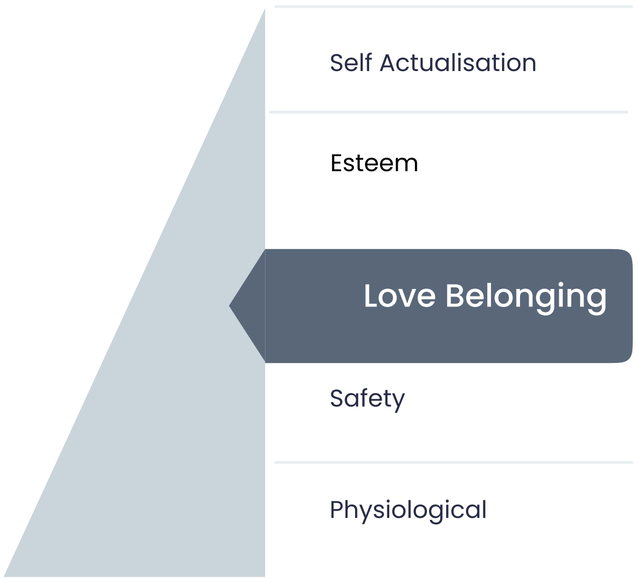
54

Gurgaon

Nursing

Retired

Maslow Pyramid



Description

Uma is a retired individual who enjoys using social media to stay connected with friends and family and learn about new things. She is not as tech-savvy as her younger counterparts, but she is still active on social media. Susan is concerned about the issue of misinformation on social media, but she feels that she does not have the tools or knowledge to effectively identify and avoid it.

@ Goals

- To learn more about misinformation and how to identify it.
- To find reliable and trustworthy sources of information.
- To feel more confident about the information she shares on social media.

Frustrations

- The overwhelming amount of information on social media, which makes it difficult to distinguish between real and fake news.
- The lack of clear and easy-to- understand information about misinformation.
- The feeling Of being overwhelmed and unsure of what to believe.

Empathy Map

Says

- I don't always trust the information I see on social media."
- I wish there were more tools to help me identify misinformation

Thinks

- I need to be more critical of the information I see online
- I'm not always sure how to identify misinformation.
- I want to be a more responsible social media user

Does

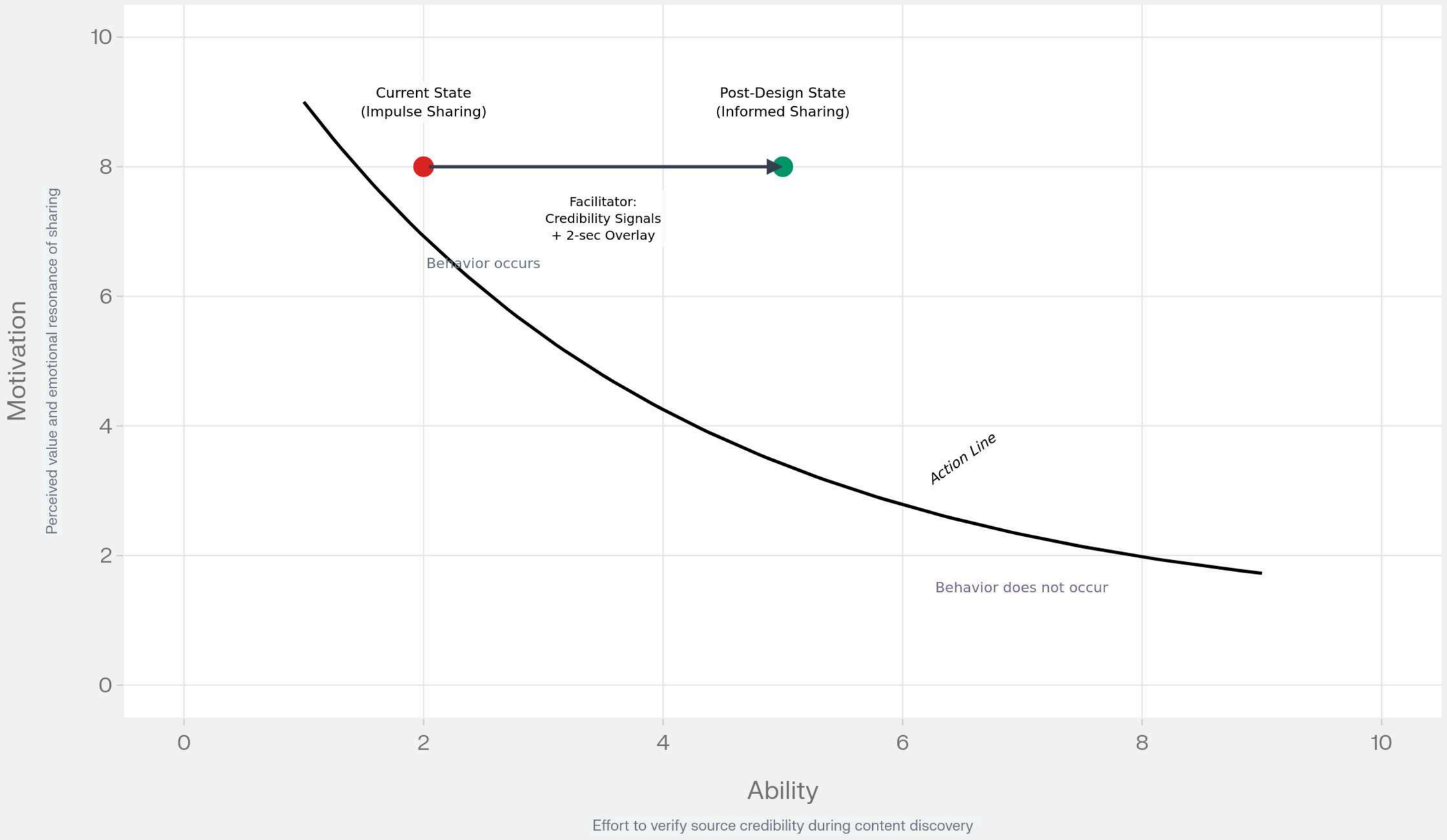
- Scrolls through his social media newsfeed.
- Shares content with friends and family.
- Reports misinformation when he sees it

Feels

- Overwhelmed by the amount of information on social media"
- Concerned about the spread of misinformation
- Frustrated by the difficulty of identifying misinformation

Behavioral Analysis

Fogg Behavior Model: Misinformation Sharing on Instagram



B=M×A×P

B = Behavior (verify before sharing) | M = Motivation | A = Ability | P = Prompt

CURRENT STATE (BEFORE INTERVENTION)

Motivation (M)

9.0/10 ✓

Users care about accuracy. 87% said "I want to share responsibly." Pain of regret is high.

Ability (A)

1.5/10 ✗

Verification takes 2–5 min. No integrated tools. 64% say "too complicated." BOTTLENECK.

Prompt (P)

8.0/10 ⚠

Fact-check labels visible. Community notes exist. But prompts don't translate to action.

Current Behavior

B = 108

9 × 1.5 × 8 = Impulsive, unverified sharing dominates.

DESIGN INTERVENTION: INCREASE ABILITY

Strategy: Reduce verification time from 2–5 min → <10 sec via one-click fact-check, credibility badges, and pre-share modal.

Ability (A)

1.5 → 6.5

+5.0 ⭐ Primary Impact

Projected Behavior

108 → 557

+416% Improvement

 # Ability is the gap (1.5/10); this is where design can make the biggest impact

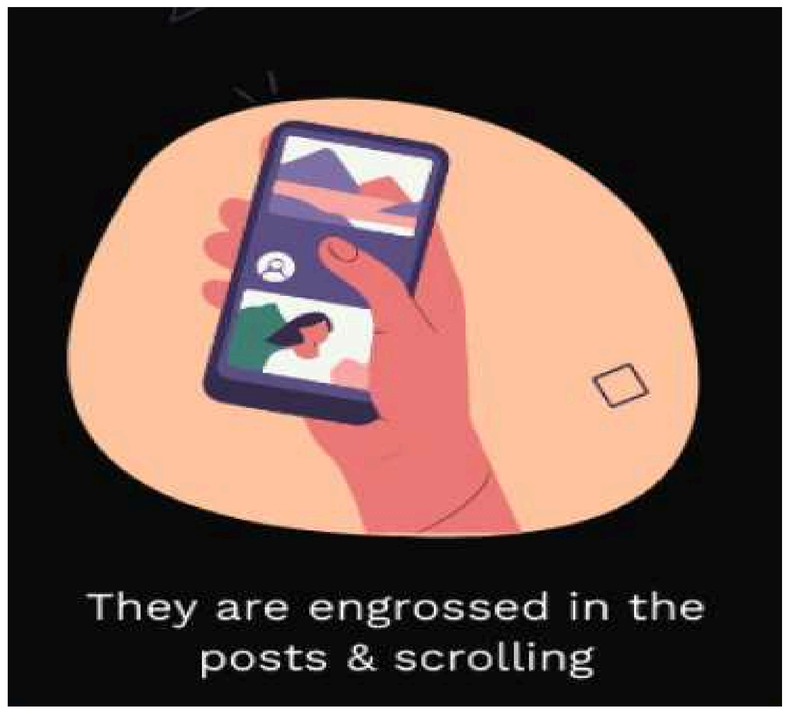
User Journey Map

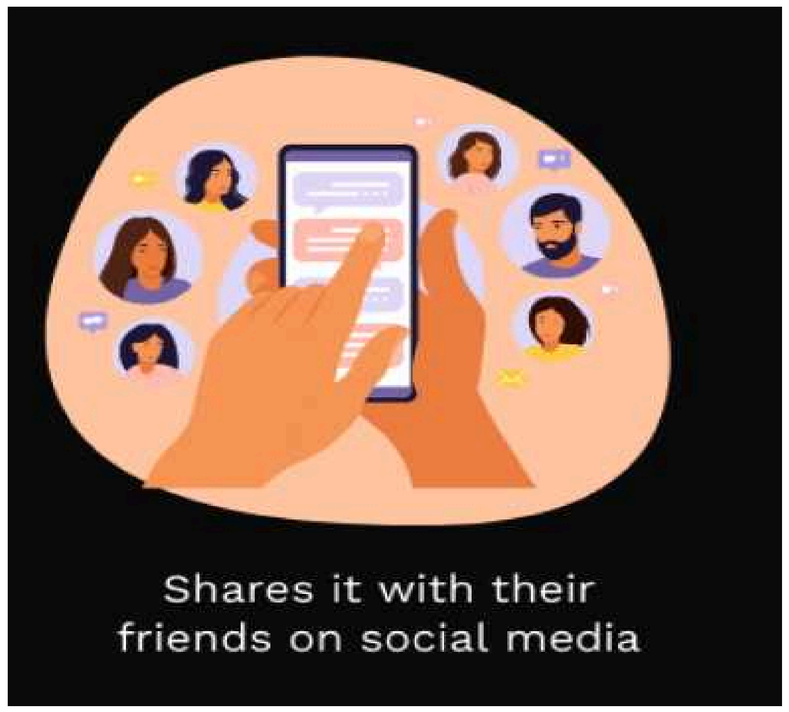
Stage	Actions & Tasks	User Emotions	Pain Points	Improvement Opportunities	Design Implications
STAGE 1 **Signing Up**	• Create an account and verify identity. • Provide personal information (email, DOB, handle). • Set basic privacy and content preferences. • Follow initial recommended accounts.	**Excited, Curious** 😀 Optimistic about joining, slightly cautious about data privacy.	• Overwhelming multi-step onboarding flow. • Confusing or hidden privacy settings. • Unclear consent language for data use. • Difficulty deciding whom to follow at the start.	• Streamlined, progressive onboarding with clear stages. • Plain-language privacy explanations and previews. • Smarter follow recommendations based on user intent. • Early trust-building moments (security hints, control toggles).	• Limit signup to the essentials; defer non-critical steps. • Introduce a privacy walkthrough with visual sliders. • Show "example feed preview" before commit. • Surface security and privacy badges prominently.
STAGE 2 **Browsing Content**	• Scroll through feed and Explore tab. • Read captions and comments on posts. • Watch short videos and stories. • Save or like posts for later.	**Entertained, Overwhelmed** Enjoying novelty but starting to feel saturated and distracted.	• Information overload from infinite scroll. • Hard to distinguish credible sources from random accounts. • Algorithm reinforcing similar content, creating echo chambers. • Sensational or misleading headlines dominating attention.	• Clear source and author labeling on content. • Adjustable content density and topic filters. • Feed diversity indicators and recommendations. • Fact-check labels on trending or viral posts.	• Add small, consistent credibility badges next to usernames. • Show origin (publication, expert, peer) within post metadata. • Provide "tune my feed" controls within the feed UI. • Use subtle prompts to nudge breaks after long browsing sessions.
STAGE 3 **Engaging with Content**	• Like and react to posts. • Comment and reply in threads. • Tag friends or influencers. • Participate in heated or emotional discussions.	**Connected, Agitated** Feels part of a community while also emotionally activated by polarizing content.	• Echo chamber dynamics amplify one-sided narratives. • Exposure to hostile or toxic comment sections. • Social pressure to align with group opinions. • Limited visibility into context or opposing views.	• Context panels explaining topic background. • Counter-argument surfacing and alternative viewpoints. • Prominent but respectful reporting and moderation tools. • Emotional resilience and digital well-being tips.	• Highlight "context added" banners on sensitive posts. • Show a "see other perspectives" CTA near comment input. • Use compassionate microcopy around disagreements. • Offer "mute topic" or "hide thread" actions in one tap.
STAGE 4 – CRITICAL **Sharing Content**	• Repost articles, memes, and videos to feed or stories. • Write captions expressing strong opinions. • Forward content directly via DMs or group chats. • Share based on emotional reaction rather than verification.	**Impulsive, Reactive** Confident in beliefs, motivated by speed and social validation.	• Hasty sharing without checking facts. • Confirmation bias reinforces existing viewpoints. • No friction or prompts before spreading questionable content. • Regret after realizing content may be misleading.	• Pre-share "pause and reflect" prompts for risky content. • Quick access to fact-check summaries and source ratings. • Optional delay/schedule for posts to reduce impulsivity. • Impact preview showing estimated reach and risk.	• Introduce a 2–3 second confirmation screen for flagged content. • Display clear "This content is disputed" messaging before share. • Offer a "share with context" template encouraging explanation. • Provide an easy "undo within 30 seconds" option after posting.
STAGE 5 **Finding Out Later**	• Learn that shared content was misinformation. • Receive corrections, DMs, or platform warnings. • Re-assess personal judgment and reputation. • Consider deleting or correcting the post.	**Embarrassed, Angry, Reflective** 😳 Feels shame, frustration, and the need to repair trust.	• Loss of credibility among peers and followers. • Public embarrassment and fear of being labeled "gullible." • Emotional distress with limited support. • No clear, structured way to correct mistakes.	• Gentle, non-judgmental correction flows. • Visible "corrected content" indicators instead of hard deletion. • Positive reinforcement for acknowledging errors. • Short media literacy and critical-thinking tips.	• Add a "Correct this post" CTA linked to a guided flow. • Allow adding an explanation note to previously shared content. • Introduce a badge for users who consistently correct misinfo. • Surface learning resources after a correction event.

Scenario

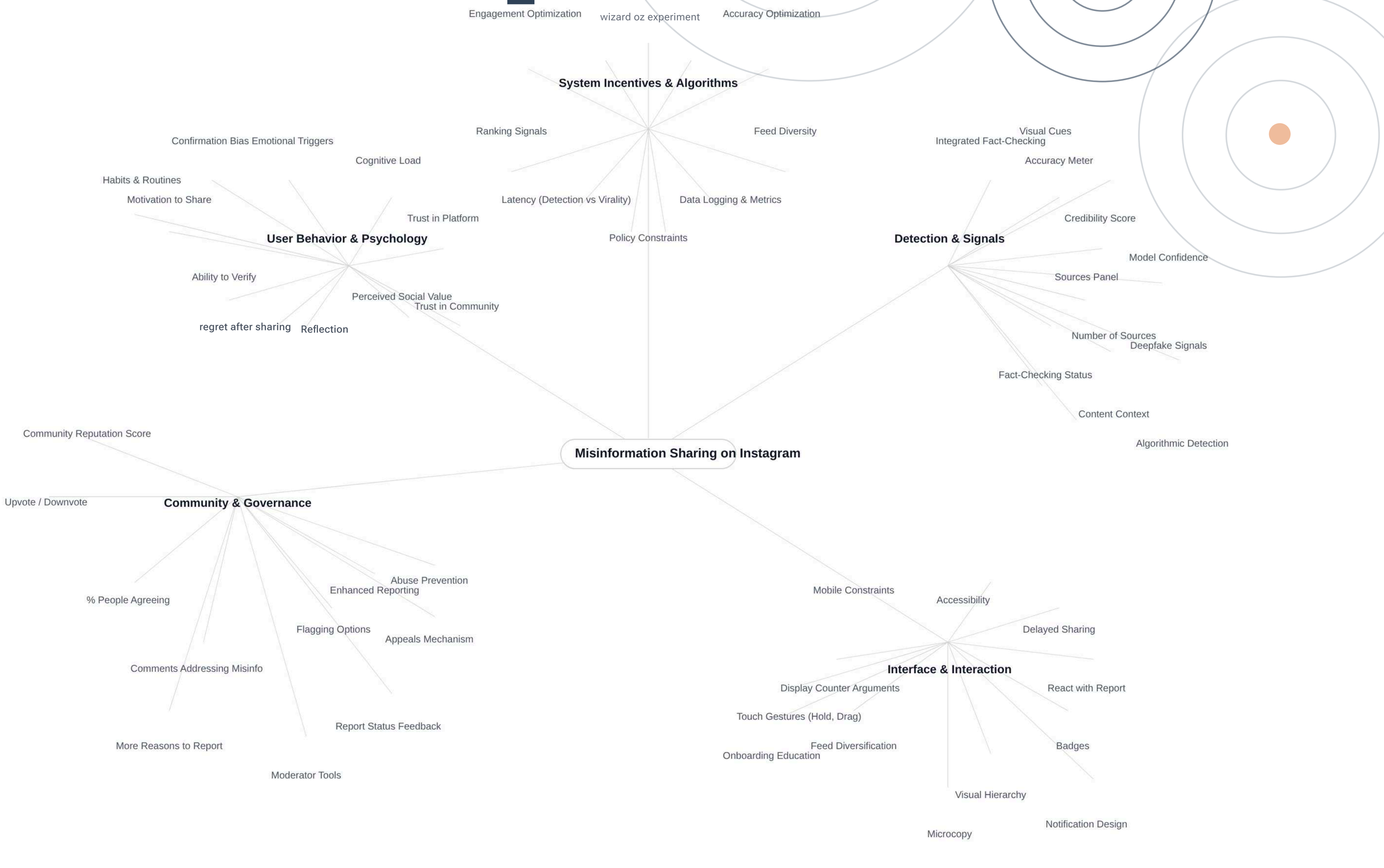
Facing Fake News

The user across potential fake news that is visually flagged with a fakeness meter. The UI is also designed to help the user engage with the content to add their view on this post



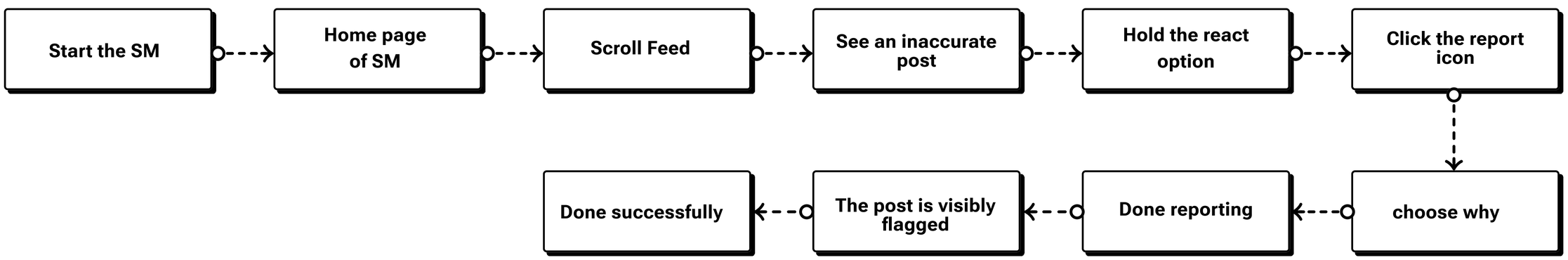


Mind Map

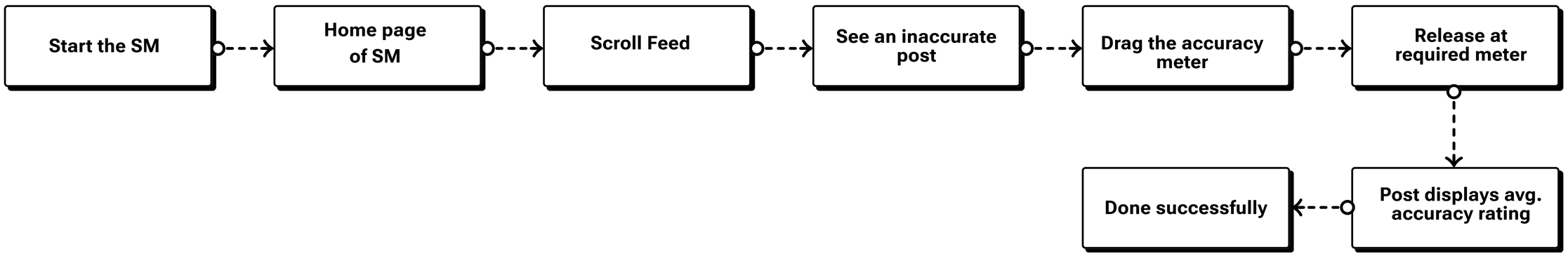

Misinformation Sharing on Instagram
System Incentives & Algorithms
Engagement Optimization
wizard oz experiment
Accuracy Optimization
Ranking Signals
Feed Diversity
Latency (Detection vs Virality)
Data Logging & Metrics
Policy Constraints
User Behavior & Psychology
Confirmation Bias Emotional Triggers
Cognitive Load
Habits & Routines
Motivation to Share
Trust in Platform
Ability to Verify
Perceived Social Value
Trust in Community
regret after sharing
Reflection
Detection & Signals
Visual Cues
Integrated Fact-Checking
Accuracy Meter
Credibility Score
Model Confidence
Sources Panel
Number of Sources
Deepfake Signals
Fact-Checking Status
Content Context
Algorithmic Detection
Community & Governance
Community Reputation Score
Upvote / Downvote
% People Agreeing
Abuse Prevention
Enhanced Reporting
Flagging Options
Appeals Mechanism
Comments Addressing Misinfo
Report Status Feedback
More Reasons to Report
Moderator Tools
Interface & Interaction
Mobile Constraints
Accessibility
Delayed Sharing
Display Counter Arguments
React with Report
Touch Gestures (Hold, Drag)
Feed Diversification
Onboarding Education
Badges
Visual Hierarchy
Notification Design
Microcopy

Task Flow

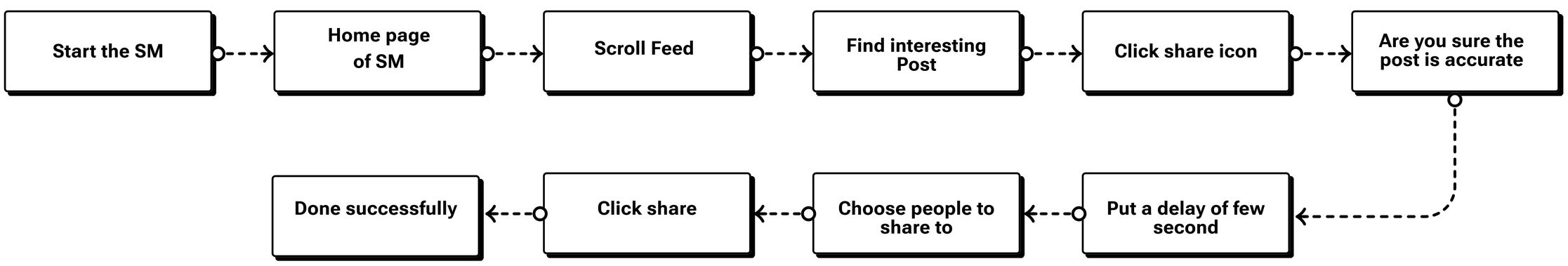
ENHANCED REPORTING



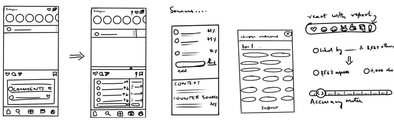
ACCURACY METER



DELAYED SHARE

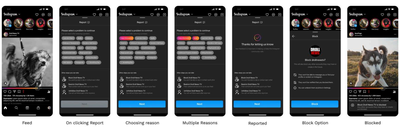


Low-Fid Prototype



High-Fid Prototype

Enhanced Reporting:

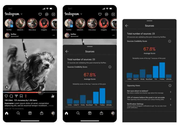


Enhance Reporting Mechanisms

Streamlining the reporting process, offering a comprehensive list of reporting options, and providing status updates on reported content.

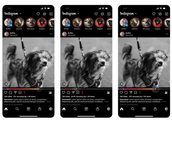
Source Visibility

View sources and their reliability. Engage in content that challenges your current belief systems to grow your critical thinking abilities.



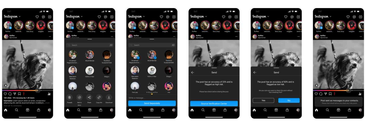
Accuracy Meter

Toggleable visible accuracy bar that is measured not just with how people rate it, but also with the sources accuracy and reliability of source.



Delayed Share

Depending on the risk evaluation framework, the post is flagged as medium, high and low risks. There are posts with no risk also, which are user's personal posts. While sharing, the level of delay is decided by the risk level.



Depending on the risk framework, the delay can be either just one more step or getting the user to fact-check before sharing depending on the risk level.

The echo chamber of AI-driven misinformation thrives on the challenges users face and the limitations of current platform defences. Our research has revealed the complexities of this landscape, highlighting user vulnerabilities and platform shortcomings.

It's a call to action for both social media platforms and UK designers. By amplifying user voices, embracing transparency, and nurturing critical thinking skills through design, we can build tools that empower users to navigate the information wilderness and break the cycle of misinformation.