Commonplace

Trust through Trickery

Caroline Sindors Vandinika Shukla Elyse Voegeli

Published on: Jan 06, 2021

DOI: <https://doi.org/10.21428/6ffd8432.af33f9c9>

License: [Creative Commons Attribution 4.0 International License \(CC-BY 4.0\)](https://creativecommons.org/licenses/by/4.0/)

Without the help and support of the following, this report wouldn't be possible: the Open Technology Fund, Tara Tarakiyee, digital HKS, Vanessa Rhinesmith, Nighat Dad, Malavika Jayaram, Tatiana Bazzichelli, Alison Carmel, Dragana Kaurin, and the Knowledge Futures Group team.

Executive Summary

This research began with an experimental hypothesis of whether trust can be measured in social networks, and consequently whether we can unpack or uncover design elements that create trust, with a focus on users facing an extreme amount of harm. Inspired by dark patterns research, we set out to uncover and test if there are design elements that can engender trust and weaponize trust in technology. Two questions follow: are certain design elements seen as trustworthy; and does such 'trustworthy' design trick users by making apps or social networks 'feel' more secure than they actually are?

The research focuses on messaging and communication apps used by journalists. As a user group, journalists face every form of harassment from domestic violence, impersonation, 'drive by' harassment (one-off harassment or harassing content posted once or twice by an individual user), to sustained harassment, doxxing, all the way to the other end of the harassment spectrum, which includes state level adverbial threats and state-level censorship. However, journalists have to be where their sources, readers, and their families are—and if they cover technology they have to use new and emerging technology. Journalists are using a wide variety of technology, apps and messaging software, even if that software is flawed in privacy or security measures. To paint a fuller picture of the landscape of harassment, the harms journalists face on platforms and the design of the products they use are two areas that should be analysed together.

For this purpose, *Trust through Trickery* is both a qualitative research study on harassment in messaging apps and social networks that journalists face, as well as a study on how design is implicated in that process, and potentially, which design elements amplify or create harassment. The paper explores if trustworthy or untrustworthy design elements could be discovered within this domain.

Design is important because the way journalists use tools, apps, and platforms, and the kind of help provided by a platform, product or app—be it in the form of policy or platform tools—need to be analyzed in relationship to harassment. Does a platform offer privacy filters, an easy and straightforward way to report harassment, any kind of comment moderation or ways to mass block? Design of these tools is important to make the tools accessible, easy to use, and easy to find for journalists to mitigate their own harassment. Algorithms that surface content are also 'designed' in that they are planned, created, and have objectives within them, around content surfacing and engagement. Thus, 'design' can be thought of as a series of choices that a service, platform, app, or product creates within their entity. Additionally, design is a necessary 'layer' in technology that makes technology, technology protocols, and policy understandable and accessible to users. In the space of online harassment though, there are no design standards, and this is similar with trust.

There is much more established research on dark patterns, which are design patterns and choices that can influence or trick users into making unintended decisions. Dark patterns¹ can be purposeful or accidental, meaning a designer or company may create a dark pattern without intentionally trying to trick a user. Dark patterns are starting to be analyzed, and named, from how they accidentally trick or confuse users. This analysis on the different kinds of patterns in dark patterns is what could start emerging in trust and harassment within design studies. In this sense, for our research, we needed to analyze what were parameters of harassment, and how design affected or created harassment with products, software and tools, and then analyze trust and its manifestations (or not) in digital design. To reemphasize, we are thinking of design in terms of writing, of how society and groups use or misuse the tool, product or service, the graphical design and user interface design, the architecture and user experience design, and the affordances and technical capabilities of the tool.

There are many varying definitions of trust, and in Silicon Valley, trust has become a buzzword and catch phrase, used alongside ‘ethical’. We looked at a variety of definitions of trust, and how different companies such as AirBnB, Facebook, and other platforms spoke about trust. We realized, based on how users relate to and use technology, that we needed a definition of trust that can work to describe the feelings users experience with their friends or families, as well as a product or service they feel like they have a choice in using. For the purposes of our research, we are using a definition of ‘trust’ coined by social psychologist Julian Rotter. Rotter defines trust as “...an expectancy held by an individual or group that the word, promise, verbal or written statement of another individual or group can be relied upon...”. Thus, we are trying to test how trust in applications is relational, i.e. does the app do what it says it will do and deliver adequately on that promise.

Our research draws from the responses of 230 designers and 81 journalists. Generally, what we found is that there are different design elements that can generate trust, from macro themes to micro themes, but the micro themes were not possible to study or uncover in this research. We conducted two workshops with twenty journalists in Mexico City affiliated with the Online News Association, ran a heuristic survey with ONA CDMX (Online News Association Ciudad de Mexico), conducted thirty one qualitative interviews with journalists in China, Hong Kong, Iran, Palestine, Malta, Guatemala, Afghanistan, the United Kingdom, Canada, the United States, Pakistan, India, Nigeria, Germany, Romania, and Mexico. Additionally, we ran two surveys, one focusing on journalists with 31 respondents focusing on harassment, and then another survey of 220 designers design, trust and dark patterns (journalists also participated in this survey). This is, to our knowledge, one of the first interdisciplinary surveys with designers and journalists to compare knowledge and awareness of design terms, dark patterns, and the processes of design in technology. This is important when establishing baselines and norms across industries and domains. How designers talk about their own work and processes, and what they think users are aware of or know, is important when unpacking technology’s effects in society and on specific user groups. An understanding of what designers know, and how they relate to specific topics and harms, and how they think about products needs to be analyzed in comparison to journalists, as well.

In our Trust and Design survey of both journalists and designers, we asked a series of questions on the design of apps, from the look and feel, to privacy and trust. Users reported that intuitive design, simplicity, clarity and ‘look and feel’ features of an app or product matter to them. We then asked users how they define trust with their favorite or most used app, and most responded with concerns over unintentional or intentional data misuse, and trusting the app not to violate that trust. While users expect privacy from the products they use, some (about ¼ of written responses) reported not trusting the app at all. Thus, consistency in messaging and consistency in data protection does facilitate trust. We then asked “how do you define trust in design?”, users responded with transparency, consistency, providing users more power or agency, and not exposing users’ personal information. These two themes—consistency and transparency—should be viewed as necessary to have trustworthy design patterns.

From our qualitative interviews with journalists in the China, Hong Kong, Iran, Palestine, Malta, Guatemala, Afghanistan, the United Kingdom, Canada, the United States, Pakistan, India, Nigeria, Germany, Romania, and Mexico, two main themes emerged: transparency and consistency. Journalists who face harassment wanted to know more about processes in platforms. Journalists, again, want transparency and consistency in platform actions. They infrequently get answers from platforms on why harassment continued, and what was considered harassment. If the harassment reports are filed on a similar topic like ‘personal threats’, to follow through on the macro trust design patterns of consistency and transparency, each of those reports should have an expected, logical outcome, and an explanation for that outcome.

The lack of transparency within platforms—especially on how they analyse harassment and choose to respond to it—breaks trust. For example, why is there neo-Nazi and anti-Semitic content repeatedly targeting journalists and users? Consistently, platforms underperform in protecting journalists, and this consistency is what makes these platforms impossible to trust. Thus, to reinforce trust in platforms, journalists demand better tools to mitigate harassment such as filters for harassment, and more consistent information flow on how their harassment cases are viewed (beyond platforms simply writing a blog post). As researchers, we understand that harassment under a policy lens can mean certain kinds of upsetting or troubling harassment is still allowed on the platform. These kinds of upsetting actions are defined in Amnesty International’s Toxic Twitter report² as ‘problematic’ or ‘abusive’. Though some forms of harassment may not ‘technically’ break the platform’s policy standards, that does not negate that the harassment is harassment. However, general users still deserve to know what or why something is considered harassment. Our findings from our interviews reveal that journalists face extreme forms of harassment or large waves of harassment. Journalists want to know and deserve to know why their reports weren’t considered harassment by the platform.

Since platforms, products, and applications repeatedly deliver poorly on the two macro themes of consistency and transparency, the micro themes of UI, color usage, and other patterns cannot be tested. When trust is broken at the macro level, it is difficult to study what can foster trust at a micro level. In essence, consistency and transparency can be thought of as trust building blocks or trust scaffolding for products meaning these

patterns can support, foster, and continue to build trust for users. It is important to translate those macro concepts into better policies and tools that make products even stronger. Trust is about giving users easy-to-understand reasons for why their harassment reports are not considered harassment, better ways to report and block content, better privacy filters, and more nuanced ways to report harassment.

Trust Through Trickery

Introduction

Trust through Trickery is a report analyzing and unpacking the multi-fold role of design and its relationship with users. Design functions beyond creating slick and usable interfaces. Design can potentially harm, trick, foster new kinds of relationships, facilitate trust and, potentially, misuse that trust. Through this report we explore what design elements (UX or UI choices) build user ‘trust’ in an app, a platform or messaging service. In examining ‘trust’ itself, we examine if ‘trust patterns’ exist broadly as a feeling that a company hopes to elicit, beyond marketing ploys and language, or as concrete, specific design choices. Our research focuses on journalists as a user group, given their status as the one group that faces every kind of harassment and harm online, from one time bullying to doxxing³ to impersonation and political or state level threats. Journalists are increasingly using a variety of different digital tools to follow breaking news stories and engage with their readers or potential sources. Their professional role and unique usage of social media or messaging apps requires them to navigate the tension between the harassment and harms they face online while being accessible for stories and tips.

This is one of the first times a research effort will have interviewed both journalists and designers about dark patterns.⁴ Journalists are the one of the few user groups that face every single form of online harassment. What platforms they use, and the negotiations they make to stay on platforms, are incredibly important to study. Designers, engineers, and platforms need to understand the threats that different groups like journalists face and how design is complicit in facilitating and mitigating these harms. Two major themes appeared in our research: consistency and transparency, which appeared not only in our trust and design survey but also in our qualitative user interviews. These two ‘themes’ need to be viewed as macro design patterns, not just best practices, not just theoretical devices, but as actual design standards.

Why Design?

Design is important because how journalists use tools, apps, and platforms, and the kind of help, be it in the form of policy or platform tools, that a platform, product or app provides need to be analyzed in relationship to harassment. Does a platform offer privacy filters, an easy and straightforward way to report harassment, any kind of comment moderation or ways to mass block? Design of these tools is important to make the tools accessible, easy to use, and easy to find for journalists to mitigate their own harassment. Algorithms, like Google Search or Facebook’s Timeline, that surface content are also ‘designed’ in that they are planned, created, and have objectives within them, around content surfacing and engagement. Thus, ‘design’ can be thought of as a series of choices that a service, platform, app or product creates within their entity. Additionally, design is a necessary ‘layer’ in technology that makes technology, technology protocols, and policy understandable and accessible to users. In the space of online harassment though, there are no formalized design standards, and this is similar with trust.

Design has a series of universal standards and protocols on how to design for web and mobile, which pulls from a methodology popularized in the 1990s called ‘human centered design.’ Human centered design, as a word and concept, was invented in 1982, coined by author Mike Cooley in his book *Architect or Bee*. Human centered design, and its methodology, was popularized by the design firm, IDEO.⁵ Human centered design is one of the reasons that a product made in Bangalore can follow a logic that is understandable to users in Singapore, or Australia, or France, for example. Even more specifically, Apple and Google/Android, have their own design methodologies and best practices. Apple’s is called Flat Design⁶ and Google/Android is called Material Design,⁷ which also is a reason that most apps from the Apple Store have a ‘sameness’ to them or a logic, where a user can download that app and not read the instructions but intuitively know how it works. This ‘intuitiveness’ is a part of human centered design, where intuitiveness and accessibility are a part of the product, as well as the logical or technical standards (e.g. is it web or mobile? What kind of technology is it using and which protocols) which is where flat design and material design are integral to the app building process. That logic is knowing that a certain button that does a certain thing goes into a certain place. Thus, design provides a structure and framework, and design has many parts from communication design and content strategy, which covers how writing is important in a product, technology or app (how things are named, how things are described), to UI and UX (how things look and where they placed), interaction design (how things move or respond), information architecture (the layout and structure of app, what flows into what), etc. In the space of online harassment though, there are no design standards, and this is similar with trust. There are some emerging best practices and thought-provoking research on design in thinking about trust, and harassment.

There is much more established design research on dark patterns, which are design patterns and choices that can influence or trick users into making unintended decisions. Dark patterns can be purposeful or accidental, meaning a designer or company may create a dark pattern without intentionally trying to trick a user. But this space of dark patterns has started to be analyzed by academic circles, with Princeton documenting over 1,000 dark patterns in e-commerce.⁸ Dark patterns are starting to be analyzed, and named, from how they accidentally trick or confuse users. This analysis and naming of different kinds of patterns within dark patterns is emerging in the studies of trust and design and harassment and design; it is important. In this sense, for our research, we needed to analyze what the patterns or vectors of harassment were, how design affected or created harassment, and then analyze trust and its manifestations (or lack thereof) in design. To reemphasize, we are thinking of design in terms of writing: how society and groups use or misuse the tool, product or service; the graphical design and user interface design; the architecture and user experience design; and the affordances and technical capabilities of the tool.

Background

Defining Trust

Trust seems to be more of a concept than a word, a bit like ethics. Societal trust is culturally rooted, and can differ culture to culture. That being said, even within culturally similar groups, communities, or groups of

people, people have seemed to inherently know about the concept of “trust” when asked but have a problem defining it. After all, trust is a part of all relationships, and it’s something experienced or learnt through time without the explicit need to define it. In a way, trust can be like breathing—we just do it without notice. But, like with a sudden coughing fit, we become keenly aware of it when something goes wrong.

For the purpose of this research, we use psychologist Julian Rotter’s definition of trust,⁹ as follows:

“...an expectation held by an individual or group that the word, promise, verbal or written statement of another individual or group can be relied upon...”

This definition of trust is applicable to both online and offline worlds, since this definition relies on expectations between different kinds of bodies (including individuals, groups, companies, collectives etc.) and the relationship within those bodies. We understand trust in contractual and relational terms. This conceptualization of trust is imperative when studying people online: how people understand and relate to each other, and their relationships, as well as how people understand and relate to the digital systems they operate within in terms of responsibility and governance.

We also use the American Press Institute’s definition, which defines trust as a set of principles that ‘aligns that with specific and actionable factors related to trust i.e. completeness, accuracy, balance, transparency, presentation and design, convenience and entertainment’.¹⁰

We also analyzed how trust is defined and used in the corporate design world. We observed that trust is often stated but not defined. The usage of ‘trust’ is either around terms like ‘transparency’ or used to define user agency. For example, AirBnB uses trust interchangeably with transparency without defining either in Medium¹¹ posts and TED¹² talks. However, books like *Calm Technology*¹³ (Case 104) do define trust in relationship to transparency, and the design firm, Projects by IF, uses trust alongside user choice and agency¹⁴.

In summary, trust is relational and we have tried to test this experience of trust through our surveys. A relational experience with an app means that it does what it says it will do and delivers adequately on that promise. Thus, we ask whether we trust our apps to perform the task it is meant for.

Impact of Trust

This report reviewed existing research on the two levels of design’s impact. Firstly, at the pure design level of the user interface and user experience which could uplift or obscure privacy and security patterns or features (for example, GDPR notifications among others). Secondly, at the level of how users interact within the designed space, for example how they communicate within an application or platform. All applications, platforms and software are designed and at the same time operate as designed spaces. Much like an architect designing a house, technologists and designers apply care and precision to platforms, applications and software. For example, designers and technologists make specific choices in what button goes where, where

the data is stored, and what a user can see. Designers may misstep and mis-design, either on accident or on purpose. Thus, it is critical to emphasize that technology is a planned space, and users can only conduct specific actions that are designed and allowed by the software, application or platform they are using. In the way that a house doesn't accidentally have X number or X numbering of growing rooms, technology allows users to engage in actions that the code in the technology has defined. Users may add their own interpretation or meaning on top of this structure, but the technology still has a confined structure to it.

These online designed spaces are closely interlinked with the harms—intentional or unintentional—within platforms or caused by platforms. These harms that users face specifically on platforms and communications applications could include online harassment, dark patterns, data misuse, security breaches, data breaches and violations, any of which could directly cause a loss of user trust in platforms.

Harassment of Journalists

While examining the design of apps, this report also conducts an in-depth analysis of harassment faced by journalists when online in social networks or using these apps. The tension of exposing themselves to danger while being available to sources, highlights the specific dangers journalists face when using technology. An acknowledged, and at times, daily occurrence in the lives of journalists, harassment against journalists can take the form of duress, stress, emotional harm and even physical harm such as murder and assault .

A study done in 2018 by the TrollBusters and the International Women's Media Foundation looks at the hostile environment harassment creates online, and how online harassment can be a silencing and censoring mechanism against journalists. As cited in this report, "Violence continues to plague journalists around the globe, including murder and assault, online harassment in the form of rape and death threats and other types of intimidation, increased surveillance, impersonation and other digital threats." Their report surveys 597 women journalists and media employees who worked across a broad media environment. Their report found, "nearly 2 out of 3 respondents said they'd been threatened or harassed online at least once — slightly more than reported physical threats or harassment. Of those, approximately 40 percent said they avoided reporting certain stories as a result of online harassment." Additionally, "Many journalists report having either abandoned their pursuit of specific stories or having difficulties with their sources as a result of the threats and abuse. Younger journalists with fewer years in the profession were also targeted; some considered leaving the profession entirely."¹⁵

In a report done by Amnesty International on female journalists and politicians across the US and the UK, their "Toxic Twitter" report found that black women face an extraordinary amount of harassment, being 84% more likely to be mentioned in abusive or problematic tweets than white women. Additionally, the report found that "7.1% of tweets sent to the women in the study were 'problematic' or 'abusive'. This amounts to 1.1 million tweets mentioning 778 women across the year, or one every 30 seconds. Women of colour, (black, Asian, Latinx and mixed-race women) were 34% more likely to be mentioned in abusive or problematic tweets than

white women. Online abuse targets women from across the political spectrum - politicians and journalists faced similar levels of online abuse and we observed both liberals and conservatives alike, as well as left and right leaning media organizations, were targeted.”

The Amnesty International “Toxic Twitter” report highlights a specific tension we noticed in our surveying and interviewing of journalists- the tension between enjoying a platform that also exposes users to harm. Imani Gandy, interviewed in the “Toxic Twitter” report said, “I think Twitter has become the new public square. I’ve found Twitter to be a really good platform for people who normally don’t have as much of a say in the political process. I’m talking primarily young people and people of colour”.¹⁶

Our study acknowledges that harassment is also a systemic issue within platforms and applications. Thus, harassment can be thought of as an unseemly and dangerous ‘feature’ of using these applications. Through the effort of analyzing design’s harmful impacts, this report seeks to shift the burden of responsibility of mitigating harassment and creating safe spaces from victims to platforms.

In a report by the Committee to Protect Journalists, when interviewing female journalists for English speaking newspapers up to the Indian National elections in 2019, they found that all “described online harassment as endemic to their work; while some said they felt the election had driven an increase in social media messages seeking to threaten, abuse, or discredit them, others said they viewed such negative engagement as routine and inevitable.” Another journalist, Dhanya Rajendran, when interviewed by CPJ, said “I’m forced to be [on Twitter], but I’m entirely dreading the experience,” Rajendran continued, saying she received thousands of harassment tweets since making a small comment about a movie in 2017. ““You know the backlash that’s going to come, and you think, is it even worth it? I now rarely voice my opinion on news,’ she said”.¹⁷

In our own surveys and interviews, we found similar feedback to design flaws and deeper questions users had about how the design works and how companies mitigate harassment and hate speech, including journalists interviewed in India and Pakistan.

Dark Patterns

Our research looked at previous reports done on malicious design, dark patterns, and parliament releases on Facebook’s data and advertising to create background on data privacy issues, as well as how the platforms (Facebook, in particular) discuss and design users, their privacy, their data, and their interests. This includes analyzing how platforms purposefully trick users or engage in acts that could potentially mislead.

Dark patterns are defined as “are tricks used in websites and apps that make you do things that you didn’t mean to, like buying or signing up for something”.¹⁸ They are malicious design patterns that can unintentionally or intentionally violate user trust and trick them into decisions they maybe wouldn’t normally make. Dark patterns have been documented in a myriad of the messaging apps in this study, from Facebook to LinkedIn. Often these design decisions can seem small, but have big repercussions for users’ safety. For example, Facebook has

been documented to set privacy features as ‘off’ by default, requiring users to actively opt-in to a safer system¹⁹.

Darkpatterns.org was created to call out specific patterns and companies for engaging in this type of profit-first, user-last design. One of their dark patterns ‘Privacy Zuckering’ is specifically named for Facebook’s notoriously tricking users into sharing more information about themselves than they would want to.

A 2018 report from the Norwegian Consumer Council, Forbrukerrådet, titled ‘Deceived by Design’ documents many of the deceptive patterns used in Facebook, Google, and Windows 10. The study found that “the combination of privacy intrusive defaults and the use of dark patterns, nudge users of Facebook and Google, and to a lesser degree Windows 10, toward the least privacy friendly options to a degree that authors consider unethical.” This kind of design is problematic for any user, but especially problematic for journalists who are already at higher risk of harassment and whose privacy can even be a life or death matter.

A growing movement in design recognizes that designers should focus on creating transparency with users, rather than manipulating them. Ultimately, good design should have the user’s best interest at heart, and not the company’s. Companies like LinkedIn have even faced lawsuits over their deceptive design patterns.²⁰

In a medium post on popular UX blog "UX Collective", product designer Arushi Jaiswal, summarizes old and new dark patterns. In our appendix, we include a list of various dark patterns. Jaiswal lists things like "misdirection, friend spam and forced continuity." Forced Continuity is "a dark pattern in which the user signs up for a free trial but has to enter their credit card details. When the trial ends, they start getting charged. There’s no opportunity to opt out, no reminder, and no easy way to cancel the automatic charging of their credit card”.²¹ The piece emphasizes that designers must take responsibility for the design decisions they make, and how these can harm users- including by violating their trust.

Categories of dark patterns (and exploitative nudges) chosen for analysis:

Default settings

- Default settings are often sticky, defaulting to least privacy friendly option is unethical
- Ask whether tailored ads are part of the core functions of the services or note, service providers should let users choose how personal data is used to serve tailored ads/experiences.
- E.g. Facebook’s GDPR popup requires users to turn off ads, accept and continuing automatically turns it on, similarly for Google have to go privacy dashboard to disable it
- Taken together, Facebook’s combination of privacy-intrusive hidden settings, and confusing wording obscuring what the “Accept”-button actually does, constitutes a dark pattern.

Ease framing

- Making progress toward the alternatives a long and arduous process can be an effective dark pattern. E.g. 41 shades of blue tested by Google
- The contrast of blue buttons for accepting, and dull grey to adjust settings away from the default, is an example of design intended to nudge users by making the “intended” choice more salient.
- E.g. Facebook: “easy road” consisted of four clicks to get through the process, which entailed accepting personalised ads from third parties and the use of face recognition. In contrast, users who wanted to limit data collection and use had to go through 13 clicks.
- Another example of dark pattern: Focusing on the positive aspects of one choice, while glossing over any potentially negative aspects, will incline many users to comply with the service provider’s wishes.
- E.g. face recognition feature in Facebook: Framing and wording nudged users towards a choice by presenting the alternative as ethically questionable or risky

Rewards and punishments

- In order to entice users to make certain choices, a common nudging strategy is to use incentives to reward the “correct” choice (e.g. extra functionality or better service), and punish choices that the service provider deems undesirable.
- Through warning users with account deletion or the loss of functionality if they decline or opt out, Facebook and Google are nudging users towards accepting.

Forced action

- All three services put pressure on the user to complete the settings review at a time determined by the service provider, without a clear option to postpone the process.

Illusion of control

- Control paradox: Studies have indicated that users who perceive that they are given more control, are also susceptible to take more risks when disclosing sensitive information. If users are only given an illusion of control, this can be considered a dark pattern used to manipulate users.
- By giving users an overwhelming amount of granular choices to micromanage, Google has designed a privacy dashboard that, according to our analysis, actually discourages users from changing or taking control of the settings or deleting bulks of data. Simultaneously, the presence and claims of complete user control may incentivise users to share more personal data.

Public Trust in Technology

This is an interesting time to start analyzing and unpacking trust. There’s a lot of disparate research out there on trust, with a particular focus on misinformation and disinformation in platforms. While this particular topic is adjacent to our study on trust and design matters, it was still important for our initial research to look at trust

and platforms in relationship to information, misinformation, and disinformation. According to a Pew Research report from May 2018²², Americans trust tech but feel conflicted. As noted by Pew, "74% of Americans say major technology companies and their products and services have had more of a positive than a negative impact on their own lives. And a slightly smaller majority of Americans (63%) think the impact of these companies on society as a whole has been more good than bad. However, when presented with several statements that might describe these firms, a 65% majority of Americans feel the statement 'they often fail to anticipate how their products and services will impact society' describes them well – while just 24% think these firms 'do enough to protect the personal data of their users...'. Meanwhile, a minority of Americans think these companies can be trusted to do the right thing just about always (3%) or most of the time (25%), and roughly half the public (51%) thinks they should be regulated more than they are now."

Methodology

Our findings are drawn from a mixture of qualitative and quantitative methods. Our study began with an initial cross-disciplinary literature review, including research from social psychology, computer science, economics, design, and tech justice. Based on current understandings of trust psychology in design, we developed interview questions for in-depth, qualitative interviews with designers and a quantitative survey for designers and journalists.

In the initial phase of our research, we conducted 10 qualitative interviews with designers, focused on their understanding of trust and how they design for trust in their professions. In a mix of focus groups and one on one interviews with these designers, we asked broad questions like "How do you define trust?" and more specific questions like "Can you recount a time in which you thought about trustworthiness manifesting in a product?". These interviews were used to ground our quantitative survey and research direction through a design lens. Common themes that emerged from the design interviews were around user choice, control, transparency and expectation-matching.

Based on these design interviews, we further further refined both qualitative interviews. The survey questions ranged in type from multiple choice to open answer. We asked users which communication apps they use in their professional and personal lives, and then asked specific questions about their most used apps like "Does the look and feel of [app] matter to you beyond functionality?" and "Think about the last time you updated your [app]. How much control did you have over the process?" and "When was the last time you checked the privacy settings for [app]?". For the full text of the surveys, please see the appendix.

We recorded 220 responses to the design survey and 31 responses to the journalist survey. To expand the findings from the journalist survey, we also conducted 31 in-depth qualitative interviews with journalists from China, Hong Kong, Iran, Palestine, Malta, Guatemala, Afghanistan, the United Kingdom, Canada, the United States, Pakistan, India, Nigeria, Germany, Romania, and Mexico about their experiences and held two stakeholder workshop in Mexico City with ONA, with nearly 20 attendees.

After an initial analysis of the quantitative and qualitative interviews, we began to see themes emerging around content moderation on platforms to perform safety. As a pilot test of these ideas, we ran a heuristics survey with a stakeholder group in Mexico and recorded 9 responses. This survey is not meant to be definitive or draw conclusions, but merely test the themes which emerged from our more indepth research.

All told, our research therefore draws from the responses of 230 designers and 81 journalists.

Findings

Summary of Quantitative Surveys

Trust Survey of UX Designers

- Demographics: gender, location

DEMOGRAPHIC DATA

Q. How old are you?

ANALYSIS

Q. What features of your most used app do you like?

Q. Do you trust your most used app? In this survey, we are using trust from the definition coined by social psychologist Julian Rotter. Rotter defines trust as "...an expectancy held by an individual or group

that the word, promise, verbal or written statement of another individual or group can be relied upon..."

Q. Do you know what Dark Patterns are?

Q. How would you define trust in design?

Q. How do you define a dark pattern?

[Design Survey](#)

Trust Survey of Journalists

[Journalist Survey](#)

Q. On which apps have you experienced online harassment? (Select all that apply.)

In our surveys, we asked a series of qualitative, write-in answers to better understand our users. In the Trust and Design survey that surveyed both journalists and designers, we asked a series of questions on the design of

apps, from the look and feel, to privacy and trust. When asked “what about the look and feel [of an app or product] matters to you?” users responded with intuitive design, simplicity, and clarity as features that reduce confusion and are pleasurable for users. We then asked users how they define trust with their favorite or most used app, most responded with concerns over data or unintentional messages, and trusting the app to not violate that trust. Meaning, users do expect privacy from the products they use, but some (about $\frac{1}{3}$ of written responses) mentioned not trusting the app at all. In particular, one mentioned trusting an app if “it hasn’t had any scandals like Facebook.” Thus, consistency in messaging and consistency in data protection and users information does facilitate trust. We then asked “how do you define trust in design?”. Users responded with transparency, consistency, providing users more power or agency, and not exposing users personal information. With transparency and consistency, this user’s response best summed up both, “For instance, when I tried to ‘pause’ Instagram but the button didn’t work, the only button that did work was ‘remove account.’ In that situation I felt like I didn’t trust Instagram.” The transparency of what the button should do was not delivered on, and the fact there was a bug was inconsistent with the expectations of a platform working. Lastly, we asked users if they knew what a dark pattern was (60% did) and then we asked them to define a dark pattern. Most users mentioned trickery, misleading design and/or persuasive design, or pushing users to do things they didn’t intend to do, like click on ads, or ignore policies, or spend money.

It’s important to note how these insights can be translated into design. The way that platforms can build trust—in consistency and transparency, with data protection, with consistent design, and with disclosures—platforms and products can also persuade users to do things against their will, such as dark patterns. Insights and analysis can lead to naming of design patterns, which makes it easier to call out malicious design choices. Dark patterns, for example, are starting to be named. In this case, these takeaways of what designers and journalists think are important to analyze.

It’s important to note, though, even best practices can have unintentional consequences or can be ‘gamed’. In our survey, users cite simplicity as a nice to have but simplicity can be misused. However, if designing for simplicity, where do new features lie and how difficult are they to find? Where are features arranged? In terms of privacy and new moderation style tools, these features are often buried in other UI. A simple interface can lead to confusing placement of key features- and we discuss this further down in the report with a specific example with Twitter and keyword blocking.

Summary of Qualitative Surveys

Journalist Interviews

Nearly every journalist interviewed (including China, Iran, and Palestine) use Facebook, Twitter, WhatsApp and Instagram. Just around half interviewed also use Telegram. The big tech apps are quite big, and have a universal reach. Nearly no one has separate accounts on one platform for the personal and the private (meaning having a public twitter account for professional relations and then a private Twitter account for personal

reasons). Generally, journalists use Twitter as the professional, public account, and platforms like Instagram and Facebook for more private accounts. Twitter is generally the most used account listed by journalists but hesitated in using the word ‘favorite’ to describe it. Journalists often would caveat by saying Twitter was a difficult place or with its problems, but use it mainly for work. Instagram was often mentioned to be the favorite, and only a small number of our journalists interviewed faced harassment on Instagram (this could generally be because journalists use Instagram in a more private and personal way, and less like Twitter, where they post more frequently and publicly).

Our interviewees have faced a wide variety of harassment, from name calling, gendered or racial slurs, to doxxing and or repeated doxxing, death threats, rape threats, and others. In the next few paragraphs, we will be covering the harms and harassment that our interviewees and survey takers have faced.

Journalists expressed frustration and confusion over how platforms responded to harassing content, in particular white supremacy. Additionally, journalists highlighted needs and hopes from the platforms, such as wanting better understanding around how certain kinds of hate speech and content ‘were allowed’ to be on the platform, of wanting better or faster responses to harassment, and wanting platforms to more deeply understand the pressures that journalists face. From our user interviews, we crafted two workshops and a heuristic survey to run with ONA CDMX (Online News Association Ciudad de México). We then synthesized the journalists’ responses which led us to the two main themes of ‘trust patterns’ or rather, ‘lack of trust patterns’ that emerged of consistency and transparency.

ONA Heuristics Survey

In October 2019 and February 2020, we ran two workshops and then in January 2020, we ran a heuristics survey with a stakeholder group, ONA Ciudad de México, on potential changes platforms could make to ensure journalists’ safety. These heuristic suggestions came from the qualitative user interviews, as well as the first ONA Ciudad de México workshops. Nine participants took our survey, which is about half of the nearly twenty participants who attended our first workshop, with ten attending the second workshop.

All of the journalists surveyed use Facebook, Twitter, WhatsApp, Instagram, email clients (correo electrónico) and a majority used Telegram and Youtube. All except one of the journalists surveyed in the heuristic survey has faced harassment on one or many of the above platforms, with Facebook and Twitter being the platforms where they faced the most harassment.

¿Qué plataformas usas?

0 / 9 correct responses

All except one of the journalists surveyed has faced harassment on one or many of the above platforms, with **Facebook and Twitter being the platforms where they faced the most harassment.**

We created a set of heuristics that focus on privacy and safety for journalists, that pulls directly from key scenarios that were highlighted in our qualitative interviews. These key scenarios are kinds of attack vectors journalists face when existing online. For example, journalists need to be accessible to the public to follow up on breaking news and be available to sources, but by having a lack of content moderation tools at their disposal, they cannot self mitigate harassment that they face. Thus, journalists have asked for solutions like algorithms to find or suppress anti-semitic (or generally harassing) content, to being able to just have ‘better’ tools. Journalists have mentioned ways to report entire hashtags that are created to harm victims (e.g. not just selecting isolated content using the hashtag, but flag the entire hashtag for a platform to investigate). Under better tools, our team decided to explore more what better could mean for journalists. Journalists wanted better reporting tools, easier ways to block or mute content and hashtags. Below, screenshots show where the muting and blocking keyword functions lie in Twitter- it’s buried under three steps, and is hard to find.

Having listened to the user interviews, we realized that journalists were grappling with a specifically tenuous and stressful situation- needing to be public to do research and interact with sources but need some of the safety and affordances of being private. **As design researchers, we asked a series of heuristic questions that focus specifically on potential tools that could be made.**

Below, we will translate and share the results. Bear in mind, No means no, Sí means yes, and No estoy segura/o means I’m not sure.

A few highlights of the responses are shown below.

The ability to report a trending hashtag

La posibilidad de reportar las etiquetas (hashtags) que estén en tendencia.

9 responses

The ability to improve harassment reporting to the platform by selecting many different posts and content and add those urls to the ongoing harassment report:

La posibilidad de mejorar los informes de acoso al seleccionar diferentes publicaciones o contenido y poder agregar esas URL al informe de acoso en curso:

9 responses

*For example, not just selecting 5 tweets surfaced by twitter but allowing a user to search for relevant tweets and add them

The ability to edit a harassment report even after it has been submitted.

La posibilidad de editar un informe de acoso incluso después de que se haya enviado.

9 responses

The ability to request or see more information about who the content moderator is

La posibilidad de solicitar o ver más información sobre quién es el moderador de contenido.

9 responses

The ability to request more information about how content moderators made their decision(s)?

La posibilidad de solicitar más información a los moderadores de contenido sobre cómo toman las decisiones.

9 responses

The following heuristic selections were 100% in favor by our users:

- The ability to ‘contest’ or challenge a content moderator’s decision on your harassment report? (meaning, when your report comes back and it says “we viewed this as not breaking our guidelines” you would be able to say, I want to reopen this case and have another person look at it).
- Turning off the ability to share or retweet your content (while not going private)
- The ability to have more transparency into how a platform's algorithms understand or filter out ‘bad’ content. This could include explanations, examples, and updates/changes made by the platform

Observations

Almost all journalists interviewed have been doxxed in one way or another- either, repeated, systematic doxxing due to who they are or the ‘beat’ or topic they cover, or in relationship to one article they wrote which ‘kickstarted’ harassment (often, this article has gone viral or popular). Nearly all the journalists interviewed, including ones in Palestine, Iran, and China, are on Facebook, Twitter, and WhatsApp.

Messaging apps played in an important role in newsrooms and for journalists. A fair amount of non-global north and non-US focused groups have commenting sections (including Germany) and have WhatsApp or telegram groups for their publishing entity/journalism site. Some of the users interviewed are the moderators for these sites. However, Telegram’s security is often called into question. In an interview in 2019, Mahsa Alimardani, an Iran Internet researcher working with Oxford Internet Institute and Article 19, helped better contextualize how Telegram’s design can provide security via the design of the chat app, even if it’s software isn’t as secure as Signal’s. Alimardani explains, "If you spend time going through case files of arrests in Iran, the majority of the time you find the arrests are based on public social media posts or evidence gathered from seized devices. How good the encryption on the chat app you used won't be of much use in this scenario.

Destroyed logs of chats will." Alimardani is highlighting threat modeling- if a user isn't concerned with message interception, but is concerned about the device being confiscated, then a security protocol isn't that helpful but an easy to use and easy to find UI choice to delete an entire conversation is.

"But Telegram, you have to use the V.P.N. to use Telegram. The facts of inconvenience just makes you feel like the Telegram is more secure, like the government has no way to control it, so they just banned it. It just makes me feel like this app..maybe is more secure than others."- a user interviewed from China

Etienne Maynier, a technologist with Amnesty International, said in a 2019 interview "Telegram has a reputation of being very secure and a part of that is clearly marketing. For instance, Telegram doesn't implement end to end encryption by default, and even when it is enabled in secret chats, it's badly implemented. Cryptography experts have reviewed it and it is considered way less secure than Signal end to end encryption." However, there are designed elements of Telegram that people like and it functions much like a social network. Maynier continues, "Telegram is also very much a social network, supergroups can have up to 200 000 users and the channels allow to broadcast information to a large number of users. There is no chat application that has this wide network, similar to a social network. For instance, people in France are organizing in the coronavirus crisis on Telegram because it is easy to send an invitation link and join groups with thousands of people..."

What both Alimardani and Maynier are highlighting are designed choices inside of Telegram- the ability to have easy to access and understand deleting messages, as well as how Telegram functions and feels like a social network- the ability to add lots of people to chat, and that chat can function like a mini social network. Additionally, Telegram's usage can be explained through popularity via location, such as Telegram being a popular chat app in Iran before it was blocked and after, and then also the affordances of the tool. What we've heard throughout years of our design research practice is an interest and liking of 'whimsical' design touches like better gifs and stickers (though we had a hard time verifying this for the research report, if we had a wider user group, we could test this). For example, we've noticed users outside of Mainland China having mentioned enjoying the app Line because of stickers. The same sentiment of global north western users has been emphasized and reemphasized about Telegram with it's stickers, video features and reactions. These things are helpful and enjoyable to users.

Another user in China mentioned, "the Telegram is quite easy to navigate. It just has a really fancy platform. I'd say the interaction with Telegram is really satisfying and it's really easy to use." Lastly, another user in China emphasized that people think of Telegram as a more secure chat product but also mentioned how widely used WeChat is. Signal is not banned in China but Telegram is. This user also mentioned feedback on how a colleague had been questioned by the Chinese government over information that had only been shared on WeChat. This kind of story has seemed to happen before, with conversations being accessed by the government over WeChat.

Platform harms

General Harms

Our interviews and surveys offered some trends on the types and experience of harms faced by journalists on platforms. Every journalist we interviewed experienced some form of online harassment. In summary, the most commonly reported type of harassment include multiple instances of doxxing, targeting based on political opinion, very extreme forms of political harassment including government targeting families of journalists in Malta, and threats from large scale actors like the Taliban. The sense of safety for journalists on platforms varies on a range of some feeling a sense of safety on Whatsapp and Instagram, and feeling no sense of safety on Facebook, and many reporting Twitter as the least safe platform. Some outliers include Discord and Instagram which users find safe due to its private close communities and options to have comments filtering on. Chinese journalists that have to use a VPN to access Telegram point to an interesting false sense of security that is built by the idea that inconvenience makes the user feel Telegram is more secure or that the government has no way to control it.

The experience of harm on platforms has specific cultural contexts as well in regions like Central America and South Asia. For example, in Mexico groups censoring speech are notorious with bots or trolls censoring, attacking and doxxing public speech. Moreover, as Chinese journalists highlight, given their location they have to also deal with both Chinese media companies with no available mechanisms to report harassment on such platforms. One of our journalists highlights that there are many different thresholds of safety especially with what Westerners or North Americans consider 'safe' or 'not safe'.

Harassment on platforms has been experienced in 'peaks' and related to stories being published, with patterns repeating themselves, instead of stand-alone individual instances of harassment. This is juxtaposed with the reality that journalists cannot report a phenomena and can only report individual content, to which then platforms revert with saying it wasn't a violation of their terms and conditions. While journalists were targeted in response to their stories, harassment was frequently targeted more as an attack on the journalists ethnic, religious or gender identity and personal character.

Q9 - Have you experienced the following online?

Offensive name-calling

Doxxing i.e. is someone searched for and published private or identifying information about you on the Internet, typically with malicious intent

Purposeful embarrassment such as direct efforts to humiliate you online, or embarrass or shame you for your opinions you online, or use of embarrassing/comprising photos or videos.

Physical threats Sustained Harassment e.g. repetitive online harassment over a sustained period of time

Online stalking Threats of public stalking Sexual Harassment

Serious violent threats such as rape or death threats Any Harassment

Actions such as digging up old tweets or online content posts taken out of context

Fabricated social media posts to antagonize, harm or incite violence Other

Q10 - What has the harassment targeted?

Q15 - On which apps have you experienced online harassment? (Select all that apply)

Self Censorship

From our interviews and workshops and heuristic surveys with ONA CDMX, journalists have highlighted how inconsistent feedback or decisions from platforms on the harassment reports they file are. Four of the journalists in the United States mentioned using a tool called DeleteMe to help combat doxxing, but most journalists felt like there weren't any tools at their disposal. Most often, instead journalists would augment their own behaviors and engage in self censoring, such as deciding to not post certain content such as politics or personal details. Journalists carry a cognitive burden of engaging in threat modeling (though, most didn't refer to it as that) but of thinking about what would happen if they posted content that could potentially reveal their family members or location. One journalist had an anecdote about realizing, while in our interview, they use social media in a self censorship way- similar to walking around a city- as in being very conscious of their actions to avoid harm.

Reporting and Algorithms

Generally all mentioned reporting, though not having confidence in the reports accomplishing any kind of outcome. Often reports would come back from platforms with the reported content labeled as 'not harassment.' Journalists mentioned how distressing this was, but also confusion over the reporting and no one to glean further insights or understandings from platforms. "How and why" what is considered harassment were common themes. In this similar vein, one quarter of journalists interviewed mentioned wanting some kind of algorithmic intervention or filter on anti-semitic and or harassing content and better tools to recognize white supremacy accounts. One user from North America specifically mentioned how Twitter has done this with ISIS in lessening and suppressing ISIS content online, so why not white supremacists? Another user from the UK said, "time and time again, these things [harassment caused by white supremacists and new white supremacy accounts] happen and they don't get caught by the filters, and they say that the algorithms aren't good enough.

Now, I don't know if that's the case or not. If the algorithms aren't good enough, then why not? And, if they really aren't good enough, with all the money that someone like Facebook can throw at this problem, why aren't they good enough?"

Consistency and Transparency

Users want more knowledge on what platforms are considered to be 'bad' or harassing content, and what is not. The journalists we interviewed face varying forms of harassment, from sexism, racism, and threats, but all who have mentioned facing harassment have mentioned an inconsistency in platforms in what is considered harassment and what is not.

Thus, two main themes emerged- transparency and consistency. Journalists who face harassment wanted to know more about processes in platforms. what they wanted was transparency, and then they wanted consistency in actions from the platforms. transparency into why their harassment continued, and what was considered harassment- and infrequently do they get answers from platforms. we discovered that we actually can't name small trust patterns—yet. since the lack of transparency within platforms is so bad, of how platforms analyze harassment, of why there is neo-nazi and anti-semantic content that journalists (and users) repeatedly face. all of those types of harassment and how platforms choose to respond to it is what breaks trust. Consistently, platforms under perform to protect journalists, and this consistency is what makes these platforms impossible to trust. So to gain trust, journalists have asked for better tools to mitigate harassment such as filters for harassment, and more consistency into how their harassment cases are viewed beyond platforms just writing a blog post.

Usability and Color

In our background research of trust patterns in design, aesthetics (especially color) and usability were frequently listed as key elements of trust-building with users. In a world where 90% of an assessment for trying out a product is made by color alone²³, it is a key design decision. For example, the color blue is the most commonly used in app logos. In color theory, blue has a variety of positive associations. Dark blue is associated with stability, light blue with calm, and bright blue with energy. In surveys of colors, blue is fairly neutral across different cultures and popular across genders. Blue has become the app color of choice for many messaging apps²⁴.

While the research certainly indicates that color plays a role in building trust with users, color was not mentioned in any of our interviews or surveys. There are several explanations for this; perhaps these aspects of design are subconscious and not actively noticed by users. Or, perhaps things like consistency and user safety are more top of mind when it comes to trust. In any case, color is an important aspect of trust, but not one which we users to be actively thinking about in our research.

Recommendations

Improved tools for journalists

Throughout our surveys and user interviews, journalists often mentioned things they wish existed, as well as problems they've encountered from the different products and platforms they use. Journalists want things like easier abilities to block or mute hashtags, better harassment reporting tools, and more insights into how platforms and content moderators make decisions. Additionally, key issues journalists have faced, like a sadness and confusion over harassment report outcomes, but no way to follow up or contest a report, shows there is a need for users to better engage in what could be described as a 'circular' system with reports. Currently reports are linear, they go in a flow from start to finish, but a circular system would allow for a report to be reopened, with future dialogue. Currently, when a report is given an outcome, that report is articulated to users that the report is closed with that decided decision.

Transparency of platform features

Platforms can create transparency in their decision making, and in their processes. In our ONA Ciudad de México heuristic survey, we asked questions on topics, themes, and abilities within platforms that, arguably, technically exist. We advocate making these features easier to find. For example, one could argue that the option to block trending hashtags exists under keyword muting (currently, there is no keyword blocking). However, keyword muting involves a multistep process of clicking on "more" in the sidebar on desktop, then selecting 'content preferences' and then selecting 'muted words'. Twitter could write more quickly accessible UI to immediately block, mute, or report hashtags on desktop by implementing a shortcut that exists on hover and immediately sends the user to the correct UI page. Alternatively, when selecting a hashtag, Twitter could give an option to mute, block, report in this drop down space.

Twitter recently has added changes in the last month to better and more intuitively mute hashtags. We first conducted our UI audit in February 2020 and will compare the two different settings.

What Twitter has currently implemented as of March 25, 2020, which is a less arduous process for users. Even less arduous, Twitter still could immediately send users to the page to mute or block words or hashtags as opposed to creating a three step process. The process below:

Users select the ellipsis menu (the three dots next to the search bar)

Which leads to this pop up menu where users select search settings:

Selecting 'search settings' leads users to hiding sensitive content and or blocking and muting accounts. However, once selecting 'hide sensitive content' there isn't clear confirmation that the user has muted any words or any clear indication around what 'hiding' sensitive content means, though, still, to specifically keyword mute, a user has to find their security settings in the nav bag (explored below)

What Twitter has had previously implemented on February 14, 2020 which is a more arduous and confusing process for users:

To this page. From where, what feels intuitive for a user? The previous drop down menu said search settings when highlighting a word but it didn't say "set content preferences." Is content preferences immediately understandable to a user if they are wanting to block or mute a word?

Settings	Content preferences
@carolinesinders	Explore
Account >	Search settings >
Privacy and safety >	Trends >
Notifications >	
Content preferences >	Languages
	Recommendations >
General	Select which languages you want recommended Tweets, people, and trends to include.
Display >	Safety
Data usage >	Muted >
Accessibility >	Blocked accounts >
About Twitter >	
	Personalization and data
	Personalization and data > Allow all

Once in content preferences, users click on 'muted' under safety and that leads to muted accounts and muted words.

Settings	Muted	
@carolinesinders	Muted accounts	>
Account >	Muted words	>
Privacy and safety >		
Notifications >		
Content preferences >		
General		
Display >		
Data usage >		
Accessibility >		
About Twitter >

However, it's still important to emphasize the confusing nature of Twitter's control settings. This is the home bar and it's shown here to emphasize that 'show settings' isn't immediately readable. The home bar is one of

the consistent, universal navigation tools a user has throughout Twitter. If a setting isn't readable from this bar, it can be difficult to find. Only clicking on more, then settings and privacy, then content preferences and then muting, does it lead to space for users to keyword mute.

Combating online harassment, violence and abuse of women on platforms

Like other related and adjacent reports on journalism and harassment- like the 2018 study conducted by TrollBusters International Women's Media Foundation (IWMF)'s "Attacks and Harassment: The Impacts on Female Journalists and their Reporting"²⁵, and Amnesty International's 2019 "Toxic Twitter", we recommend increased transparency from platforms.

The IWMF report suggests that organizations establish a protocol for educating and addressing harassment and that their claims be thoroughly investigated by management, law enforcement, social media platforms and others. However, there are no such mechanisms of support for freelancers, and many news organizations lack the expertise and resources to respond effectively.

Amnesty International's Toxic Twitter²⁶ report outlines the following recommendations for Twitter to address violence and abuse against women online that are meaningful for our research on online harassment of journalists across social media messaging platforms:

- Publish comprehensive and meaningful information about the nature and levels of violence and abuse against women, as well as other groups, on the platform, and how they respond to it.
- Improve reporting mechanisms to ensure consistent application and better response to complaints of violence and abuse.
- Provide more clarity about how the platform interprets and identifies violence and abuse on the platform and how it handles reports of such abuse.
- Undertake more proactive measures in educating users and raising awareness about security and privacy features on the platform that will help women create a safer, and less toxic Twitter experience.

Moving Forward: User Groups and Role of Content Moderation

Generally, we are recommending that this study be done with two more user groups, perhaps a more general audience which faces lower or less abusive harms on platforms, and another group of users which is readily online and faces low to high forms of harm (these could include digital influencers and digital activists- those that deeply enjoy being online or need to be online for work but face harm from being online).

Additionally, we found a small data point or hypothesis when conducting our user interviews that we believe needs to be further explored, and one that seems to be underexplored in the field of online harassment. When conducting our user interviews, we realized that interviewees were discussing things that seemed related or adjacent to content moderation, specifically how a platform judges or decides an outcome on an individual report. However, none of the interviewees directly talked about content moderation, they talked about the

platform at large. We decided in one of our last interviews to directly ask if the user thought of content moderation or moderators when filing a report or saw them related at all to the reports or the decisions that the platform made. The user responded it's complicated and was not sure of an answer. In a follow-up user workshop with CDMX ONA, we also asked users their thoughts about this relationship. Our users in that group revealed that when filing reports, they don't immediately think of content moderators but view the platform as one large entity, such a "Facebook being responsible for this report." However, they are aware that moderators are involved in this process, and that moderators can have a specific responsibility in deciding cases. But when filing cases, even if the journalists are aware of this function (from the ONA journalists group user interview, journalists knew what content moderators were and what they were responsible for on platforms) they aren't quite thinking about content moderators or individuals when filing reports. We believe that more research needs to be done exploring this connection of awareness and lack of awareness of content moderation, platforms, and responsibility within online harassment. By analyzing these overlapping areas more (content moderation and online harassment), we are curious what solutions or new policies could emerge in terms of victim awareness, moderator support, victim support or pushing for more transparency from platforms on how moderators are trained to analyze harassment and what support or tools moderators have specifically related to harassment cases.

Conclusion

What facilitates or designs trust? It's repeated consistency and transparency within the tools as well as clear and consistent responses to users about the harassment they face. To create better products, platforms need to unpack what transparency means in design, and that does not simply mean open sourcing code, it means explaining what a product is doing when and why, and offering opt in versus opt out provisions. In essence, consistency and transparency can be thought of as trust building blocks or trust scaffolding for products. It is important to translate those macro concepts into better policies and tools that make products even stronger. Trust is about giving users easy to understand reasons for why their harassment reports are not considered harassment, better ways to report content and block content, better privacy filters, and more nuanced ways to report harassment. The study of micro trust patterns can be studied only once these basic, macro themes are delivered on. Users need agency to make decisions within products and they need clear and explainable options. Trust is not just telling users "we are using your data in X way," it's letting them have options. Trust needs to be clear, understandable, and consistent. Legibility, intuitiveness and privacy are incredibly important in designing for trust and cannot be glossed over. They need to be equal to the design and vice versa. Trust is about legibility and honesty or 'consistency' and 'transparency.'

Appendix

Types of Dark Patterns²⁷

- **Bait and Switch:** when a user is looking to take an action that results in a desired outcome, but instead ends up resulting in something completely unforeseen.
 - Eg: Windows 10 update, when the top right ‘x’ button was clicked, it actually initiated the update
- **Disguised Ads:** ads are disguised in the page, as if they were a part of the regular content or navigation so that users click more
- **Forced Continuity:** the user signs up for a free trial but has to enter their credit card details. When the trial ends, they start getting charged. There’s no opportunity to opt out, no reminder, and no easy way to cancel the automatic charging of their credit card.
- **Friend Spam:** when the product asks for the user’s email or social media permissions under the pretense it will be used for a desirable outcome e.g. finding friends, but then spams all their contacts in a message that claims to be from them.
- **Hidden Costs:** a user going through multiple steps to checkout and after finally getting to the last step of the checkout process, discovering some unexpected charges have appeared, e.g. delivery charges, tax, etc. as ‘hidden costs’
- **Misdirection:** when the user’s attention is guided to a specific place so they won’t notice something else that is happening.
 - Eg: setting bing as the default search engine when updating skype
- **Price Comparison Prevention:** when the retailer makes it hard for the user to compare the price of an item with another item, so they cannot make an informed decision.
- **Privacy Zuckering:** tricking the user into publicly sharing more information about them than they really intended to.
 - Eg: Facebook
- **Roach Motel:** The design makes it very easy for the user to get into a certain situation, but then makes it hard for them to get out of. For example, a subscription.
- **Sneak into Basket:** You attempt to purchase something, but somewhere in the purchasing journey the site sneaks an additional item into your basket, often through the use of an opt-out radio button or checkbox on a prior page.
- **Trick Questions:** You respond to a question, which, when glanced upon quickly appears to ask one thing, but if read carefully, asks another thing entirely.

List of Interviewees

Due to danger that journalists can face either from just online harassment to political danger from state level actors, we will only list the countries and total number of people interviewed to protect the identity of our interviewees. Profile: all 32 journalists working across radio, TV and print/online news outlets.

- Afghanistan: 1

- Guatemala: 1
- Canada: 1
- China: 2
- Hong Kong: 2
- Germany: 1
- India: 2
- Iran: 2
- Malta: 1
- Mexico: 3
- Nigeria: 1
- Pakistan: 4
- Palestine: 2
- United Kingdom: 5
- United States: 3

Survey questions

ONA SURVEY

These questions are about harassment on platforms and changes platforms could take.

1. What platforms do you use?

- a. Facebook
- b. Instagram
- c. Whatsapp
- d. Twitter
- e. Tumblr
- f. Email clients
- g. Signal
- h. Viber
- i. Telegram
- j. Slack
- k. Tik Tok
- l. Youtube
- m. Otro

2. Have you faced harassment on any of these platforms?

- a. Yes
- b. No
- c. Prefer not to say

3. Which platforms?

- a. Facebook
- b. Instagram
- c. Whatsapp
- d. Twitter
- e. Tumblr
- f. Email clients
- g. Signal
- h. Viber
- i. Telegram
- j. Slack
- k. Tik Tok
- l. Youtube

4. Think about the last time you experienced harassment on the platform. At that time, was there any tool or feature you wished existed?

Think about the last time you experienced harassment online, or imagine a scenario where you have been harassed. Do the following hypothetical features help your situation? Yes/ No/ Unsure

1. The ability to select a lot of mentions on Twitter, Instagram, or Facebook and hide them

- a. Yes
- b. No
- c. Unsure

2. The ability to block trending hashtags

- a. Yes
- b. No
- c. Unsure

3. The ability to mute trending hashtags

- a. Yes
- b. No
- c. Unsure

4. The ability to report a trending hashtag

- a. Yes
- b. No
- c. Unsure

5. The ability to improve harassment reporting to the platform by selecting many different posts and content and add those urls to the ongoing harassment report:

- a. Yes

- b. No
 - c. Unsure
6. **The ability to edit a harassment report even after it has been submitted.**
- a. Yes
 - b. No
 - c. Unsure
7. **The ability to request or see more information about who the content moderator is.**
- a. Yes
 - b. No
 - c. Unsure
8. **The ability to request more information about how content moderators made their decision(s)?**
- a. Yes
 - b. No
 - c. Unsure
9. **The ability to ‘contest’ or challenge a content moderator’s decision on your harassment report? (meaning, when your report comes back and it says “we viewed this as not breaking our guidelines” you would be able to say, I want to reopen this case and have another person look at it).**
- a. Yes
 - b. No
 - c. Unsure
10. **The ability to turn off comments or replies on any content you post on any platform.**
- a. Yes
 - b. No
 - c. Unsure
11. **Turning off the ability to share or retweet your content (while not going private).**
- a. Yes
 - b. No
 - c. Unsure
12. **The ability to make special lists (like friends lists) on more platforms (like Twitter, Instagram, etc) when posting content.**
- a. Yes
 - b. No
 - c. Unsure
13. **Having comment moderation for users tools on platforms like Instagram, Youtube, etc? (meaning if I am a Youtuber and I post a video, I can delete a harmful comment on my video).**
- a. Yes
 - b. No

- c. Unsure
14. **Having comment moderation on Facebook? (Meaning, if I write a post, I become the ‘admin’ or moderator of that post, and can delete or hide comments).**
- a. Yes
- b. No
- c. Unsure
15. **The ability to have more transparency into how platform’s algorithms understand or filter out ‘bad’ content’? (this could include explanations, examples, and updates/changes made by the platform).**
- a. Yes
- b. No
- c. Unsure
16. **What else do you wish existed? How do you think these new tools would help stymie harassment?**

Footnotes

1. *Dark Patterns*. <https://darkpatterns.org/> ↵
2. “Why Twitter Is a Toxic Place for Women”, *Amnesty International*
<https://www.amnesty.org/en/latest/research/2018/03/online-violence-against-women-chapter-1/>. ↵
3. Doxxing: to publicly identify or publish private information about (someone) especially as a form of punishment or revenge. Source: <https://www.merriam-webster.com/dictionary/dox> ↵
4. Patterns or choices in design that intentionally or unintentionally trick or nudge users into making decisions. This can be different sized buttons or different color buttons, or countdown clocks on e-commerce websites that make an item feel more scarce when it’s not. ↵
5.
<https://designthinking.ideo.com/history#:~:text=At%20IDEO%2C%20we've%20been,embrace%20of%20ambiguity%20and%20failure>. ↵
6. <https://developer.apple.com/design/human-interface-guidelines/> ↵
7. <https://material.io/design> ↵
8. <https://webtransparency.cs.princeton.edu/dark-patterns/> ↵
9. Govier, Trudy, “Social Trust and Human Communities,” McGill-Queen’s University Press, 1997 ↵
10. <https://www.americanpressinstitute.org/publications/reports/survey-research/the-meaning-of-trust-in-news/> ↵

11. <https://medium.com/airbnb-engineering/building-for-trust-503e9872bbbb> ↵
12. https://www.ted.com/talks/joe_gebbia_how_airbnb_designs_for_trust?language=en ↵
13. Amber Case, “Calm Technology”, O’Reilly Media Inc, 2015. ↵
14. *Potential Design Patterns for Open APIs in the Utilities Sector – Open APIs in the Telecoms Industry*. CHANGE ME. <https://openapis.projectsbyif.com/potential-design-patterns-for-open-apis-in-the-utilities-sector/> ↵
15. <https://www.iwmf.org/wp-content/uploads/2018/09/Attacks-and-Harassment.pdf> ↵
16. <https://decoders.amnesty.org/projects/troll-patrol> ↵
17. “Journalists Fighting Fake News during Indian Election Face Threats, Abuse.” *Committee to Protect Journalists*, 21 May 2019, <https://cpj.org/2019/05/india-elections-online-harassment-female-journalists/>. ↵
18. <https://www.darkpatterns.org/> ↵
19. Forbrukerrådet, “Deceived by Design,” 2018. <https://fil.forbrukerradet.no/wp-content/uploads/2018/06/2018-06-27-deceived-by-design-final.pdf> ↵
20. Jaiswal, Arushi. “Dark Patterns in UX: How Designers Should Be Responsible for Their Actions.” *Medium*, 15 Aug. 2018, <https://uxdesign.cc/dark-patterns-in-ux-design-7009a83b233c> ↵
21. Jaiswal, Arushi. “Dark Patterns in UX: How Designers Should Be Responsible for Their Actions.” *Medium*, 15 Aug. 2018, <https://uxdesign.cc/dark-patterns-in-ux-design-7009a83b233c>. ↵
22. NW, 1615 L. St, et al. “How Americans View Tech Companies.” *Pew Research Center: Internet, Science & Tech*, 28 June 2018, <https://www.pewresearch.org/internet/2018/06/28/public-attitudes-toward-technology-companies/>. ↵
23. “Why Facebook Is Blue: The Science of Colors in Marketing.” *Buffer Resources*, 25 Apr. 2015, <https://buffer.com/resources/the-science-of-colors-in-marketing-why-is-facebook-blue/>. ↵
24. “Blue Apps Are All Around But Blue Tones Get Less Of A Role In IOS 7’s Psychedelic Redesign.” *TechCrunch*, <https://social.techcrunch.com/2013/06/15/all-the-apps-were-blue/>. ↵
25. Attacks and Harassment: The Impact on Female Journalists and Their Reporting - *IWMF*. <https://www.iwmf.org/attacks-and-harassment/>. ↵
26. Toxic Twitter - The Solution. *Amnesty International* <https://www.amnesty.org/en/latest/research/2018/03/online-violence-against-women-chapter-8/>. ↵

27. Source: <https://uxdesign.cc/dark-patterns-in-ux-design-7009a83b233c> ↵