

Clemson University

**TigerPrints**

---

All Dissertations

Dissertations

---

7-2022

## Fair, Equitable, and Just: A Socio-technical Approach to Online Safety

Daricia Wilkinson

*Clemson University*, [dariciw@clemson.edu](mailto:dariciw@clemson.edu)

Follow this and additional works at: [https://tigerprints.clemson.edu/all\\_dissertations](https://tigerprints.clemson.edu/all_dissertations)



Part of the [Artificial Intelligence and Robotics Commons](#), [Graphics and Human Computer Interfaces Commons](#), [Human Factors Psychology Commons](#), [Information Security Commons](#), [Other Psychology Commons](#), and the [Social Justice Commons](#)

---

### Recommended Citation

Wilkinson, Daricia, "Fair, Equitable, and Just: A Socio-technical Approach to Online Safety" (2022). *All Dissertations*. 3226.

[https://tigerprints.clemson.edu/all\\_dissertations/3226](https://tigerprints.clemson.edu/all_dissertations/3226)

This Dissertation is brought to you for free and open access by the Dissertations at TigerPrints. It has been accepted for inclusion in All Dissertations by an authorized administrator of TigerPrints. For more information, please contact [kokeefe@clemson.edu](mailto:kokeefe@clemson.edu).

# FAIR, EQUITABLE, AND JUST: A SOCIO-TECHNICAL APPROACH TO ONLINE SAFETY

---

A Dissertation  
Presented to  
the Graduate School of  
Clemson University

---

In Partial Fulfillment  
of the Requirements for the Degree  
Doctor of Philosophy  
Human Centered Computing

---

by  
Daricia Wilkinson  
December 2022

---

Accepted by:  
Dr. Bart P. Knijnenburg, Committee Chair  
Dr. Kelly Caine  
Dr. Guo Freeman  
Dr. Andrew Robb  
Dr. Marten Risius

# Abstract

Socio-technical systems have been revolutionary in reshaping how people maintain relationships, learn about new opportunities, engage in meaningful discourse, and even express grief and frustrations. At the same time, these systems have been central in the proliferation of harmful behaviors online as internet users are confronted with serious and pervasive threats at alarming rates. Although researchers and companies have attempted to develop tools to mitigate threats, the perception of dominant (often Western) frameworks as the standard for the implementation of safety mechanisms fails to account for imbalances, inequalities, and injustices in non-Western civilizations like the Caribbean. Therefore, in this dissertation I adopt a holistic approach to online safety that acknowledges the complexities of harms for understudied populations specifically focusing on the Caribbean.

In this dissertation, I conduct three studies that take steps towards (1) filling in the gap of missing empirical understanding around users' perceptions of safety threats and how that is associated with their intentions to engage with supportive countermeasures, (2) understanding the gaps in current approaches to justice, and (3) developing an understanding towards the development of equitable and inclusive countermeasures.

In the first study, I conduct a region-wide survey which reveals Caribbean citizens experience high rates of exposure to online threats. Moreover, I show that by conceptually defining protective behaviors based on the threats that they address, it exposes how the perceptions of threats influences the adoption of online safety countermeasures while uncovering distinctions in perceptions depending on the type of harm.

The second study utilizes a multi-disciplinary approach to understand the state of legislative protections. Through a reflective legislative and media analysis, the study uncovered major discrepancies in the region's approach towards justice in online spaces.

Lastly, the final study incorporates the findings of these works by conducting an online experiment to test the design of justice-oriented safety countermeasures. The results provide support for the development of countermeasures that people perceive to be fair, equitable, and just.



# Acknowledgments

Without any doubts nor reservations, I would like to first give thanks and gratitude to God. I am deeply grateful to my loving family, numerous friends, and my support system who kept me afloat and motivated. Thank you to my amazing advisor Bart who would have molded me into the researcher I am. I also express much gratitude to my committee, mentors, collaborators and funding agencies who made completing this work possible.

# Table of Contents

<b>Title Page</b> . . . . .	<b>i</b>
<b>Abstract</b> . . . . .	<b>ii</b>
<b>Acknowledgments</b> . . . . .	<b>iv</b>
<b>List of Tables</b> . . . . .	<b>vii</b>
<b>List of Figures</b> . . . . .	<b>viii</b>
<b>1 Introduction</b> . . . . .	<b>1</b>
1.1 Problem Motivation . . . . .	1
1.2 Research Objectives . . . . .	2
<b>2 Related Work</b> . . . . .	<b>5</b>
2.1 Understanding Online Safety Threats . . . . .	5
2.2 Theoretical Considerations . . . . .	12
2.3 Socio-technical Perspectives . . . . .	14
2.4 Research Gaps . . . . .	15
<b>3 Exploring Safety Perceptions and the Prevalence of Threats</b> . . . . .	<b>16</b>
3.1 Overview . . . . .	16
3.2 Background . . . . .	17
3.3 Method . . . . .	24
3.4 Findings . . . . .	25
3.5 Discussion . . . . .	37
3.6 Chapter Conclusion . . . . .	43
<b>4 A Critical Reflection of Legislative Protections in the Caribbean</b> . . . . .	<b>44</b>
4.1 Legislative History . . . . .	46
4.2 Method . . . . .	47
4.3 Findings . . . . .	48
4.4 Discussion . . . . .	70
4.5 Chapter Conclusion . . . . .	73
<b>5 Investigating the Role of Fairness, Equity, and Trust in Justice-Oriented Safety Interventions</b> . . . . .	<b>74</b>
5.1 Background . . . . .	74
5.2 Hypotheses Development . . . . .	76
5.3 Method . . . . .	79
5.4 Results . . . . .	89
5.5 Discussion . . . . .	100

5.6 Conclusion . . . . .	103
<b>6 General Conclusions and Future Directions . . . . .</b>	<b>105</b>
6.1 Re-imagining Online Justice Futures . . . . .	107
<b>Appendices . . . . .</b>	<b>112</b>
A Supplemental Materials for Chapter 3 . . . . .	113
B Supplemental Materials for Chapter 5 . . . . .	121
<b>Bibliography . . . . .</b>	<b>126</b>

# List of Tables

3.1	Gender distribution per country. Countries in the second segment of the table were excluded from the analysis due to a low number of participants. . . . .	24
3.2	Description of the frequency of app usage among all participants. Note that "WhatsApp Mod" represents WhatsApp FM, GB WhatsApp or any modified version of WhatsApp*. . . . .	25
3.3	The table above describes the summary of findings related to the hypotheses testing. Items denoted by (*) signify hypotheses where coping appraisal, which comprises of self efficacy and response efficacy, is observed but self efficacy was dropped and the results reflected represent response efficacy only. . . . .	38
3.4	In the table above, we summarize the results in relation to our research question as well as offering an overview of the respective implications. . . . .	39
4.1	Overview of regional legislative protections related to the prevention and prohibition of online safety threats and data protection . . . . .	53
4.2	Regional Legislative Coverage regarding online offenses . . . . .	54
5.2	The table above describes the summary of findings related to the hypotheses testing.	98
5.1	The factors of personal characteristics with the Average Variance Extracted (AVE) and the consistency coefficients (Cronbach's $\alpha$ ), and the items per construct with item factor loadings. Removed items are colored in grey . . . . .	104
1	The survey items for the digital security model with item loading, average variance extracted, and Cronbach's alpha for each factor. Removed items are colored in grey. Trust was measured once across all models since it measured attitudes towards trustworthiness of platforms independent of harm being faced. . . . .	117
2	The survey items for the harassment model with item loading, average variance extracted, and Cronbach's alpha for each factor. Removed items are colored in grey. .	118
3	The survey items for the access and disclosure model with item loading, average variance extracted, and Cronbach's alpha for each factor. Removed items are colored in grey. . . . .	119
4	The survey items for the offline model with item loading, average variance extracted, and Cronbach's alpha for each factor. Removed items are colored in grey. . . . .	120

# List of Figures

1.1	Socio-ecological model (adapted from [126, 32]) of the multi-level individual and societal factors that influence online safety. . . . .	3
2.1	Annotated Designs: Designs varied in their level of granularity (low, moderate, high, very high) and the presentation style (app-centric versus data centric) with a total of eight designs. For each design shown, there is an identical design with the same level of granularity but different presentation style. From top to bottom: "Low granularity, app-centric presentation", "Moderate granularity, data-centric presentation", "High granularity, app-centric presentation", "Very High granularity, data-centric presentation". . . . .	8
2.2	Explanation classification scheme . . . . .	10
3.1	The figure above illustrates the proposed conceptual model for the study . . . . .	23
3.2	Reported prior victimization counts across all participants (N=551) and all observed threat categories. . . . .	26
3.3	Regional victimization trends. Numbers shown represent the standardized residuals. Color gradient corresponds to the magnitude of the discrepancy (Red is smaller than expected; Green is larger than expected) . . . . .	27
3.4	Sample-wide comparison of the perceived severity of threats across all threat categories	28
3.5	Sample-wide comparison of the perceived vulnerability of threats across all threat categories . . . . .	29
3.6	The figure above displays the SEM models for threats related to digital security . . .	30
3.7	The figure above displays the SEM models for threats related to Access and Disclosure	31
3.8	The figure above displays the SEM models for threats related to online-offline contexts	31
3.9	The figure above displays the SEM models for Harassment-related threats . . . . .	32
3.10	Marginal effects of perceived severity for online-to-offline threats . . . . .	33
3.11	Marginal effects of perceived severity for threats related to Access and Disclosure . .	34
3.12	Marginal effects of perceived severity for security threats . . . . .	34
3.13	Marginal effects of perceived severity for harassment-related threats . . . . .	35
3.14	Sample-wide comparison of the response efficacy across all protective behavior categories	36
3.15	Sample-wide comparison of behavioral intention across all protective behavior categories	37
4.1	Overview of legislative protections related to the prevention and prohibition of online safety threats. Note: The figure displays "St. Vincent" which represents St. Vincent and the Grenadines. . . . .	52
55figure.caption.33		
4.3	Overview of legislative protections related to the prevention and prohibition of online safety threats in Antigua and Barbuda. . . . .	56
4.4	Overview of legislative protections related to the prevention and prohibition of online safety threats in The Bahamas. . . . .	57

4.5	Overview of legislative protections related to the prevention and prohibition of online safety threats in Barbados. . . . .	58
4.6	Overview of legislative protections related to the prevention and prohibition of online safety threats in Belize. . . . .	59
4.7	Overview of legislative protections related to the prevention and prohibition of online safety threats in Dominica. . . . .	60
4.8	Overview of legislative protections related to the prevention and prohibition of online safety threats in Grenada. . . . .	61
4.9	Overview of legislative protections related to the prevention and prohibition of online safety threats in Guyana. . . . .	62
4.10	Overview of legislative protections related to the prevention and prohibition of online safety threats in Jamaica. . . . .	63
4.11	Overview of legislative protections related to the prevention and prohibition of online safety threats in St. Lucia. . . . .	64
4.12	Overview of legislative protections related to the prevention and prohibition of online safety threats in St. Kitts and Nevis. . . . .	65
4.13	Overview of legislative protections related to the prevention and prohibition of online safety threats in St. Vincent and the Grenadines. . . . .	66
4.14	Overview of legislative protections related to the prevention and prohibition of online safety threats in Trinidad and Tobago. . . . .	67
4.15	Results from the preliminary content analysis summarizing the total number of threats covered in the corpus organized by threat type . . . . .	69
4.16	Results from the preliminary content analysis summarizing the total harms covered in corpus by country . . . . .	69
4.17	Results from the preliminary content analysis summarizing the total threats covered in corpus by threat category . . . . .	70
5.1	Proposed conceptual model for the study . . . . .	79
5.2	An overview of the study procedure. "R" denotes that participants will be randomly assigned to one of the four conditions. . . . .	80
5.3	Example scenarios presented in the study . . . . .	82
5.4	Example non-personalized design with a justice-oriented countermeasure . . . . .	83
5.5	Example non-personalized design where an alternative justice countermeasure is not included . . . . .	84
5.6	Example personalized justice-oriented countermeasure . . . . .	85
5.7	Sample size of the proposed study with a moderate effect size . . . . .	87
5.8	Illustration of the variance in sum score for safety perceptions across all four conditions . . . . .	90
5.9	Variance in safety perceptions across all scenarios . . . . .	91
5.10	The figure above displays the SEM model. Positive relationships are depicted by black arrows versus negative relationships which are depicted by red arrows. The model is color-coded based on the type of latent variable. Green denotes manipulations, purple denotes subjective aspects around fairness, orange denotes subjective aspects about the system, and blue denotes outcomes. Pers. is an abbreviated form of personalization. J+P represents the combined effect of justice and personalization. . . . .	92
5.11	Marginal effects of the manipulations on perceptions of distributive, transformative, and procedural fairness . . . . .	94
5.12	Marginal effects on the subjective and outcome factors . . . . .	95
5.13	Country-to-Country differences among conditions for behavioral intention . . . . .	96
5.14	Country-to-Country differences among conditions for perceived equity . . . . .	97

5.15	Distribution of users' chosen countermeasure/s to respond to harm. Totals are reflective of all four conditions: C1 (Personalized and Justice-Oriented), C2 (Non-personalized and Justice-Oriented), C3 (Not Justice-Oriented but Personalized), C4 (Not Justice-Oriented and Non-personalized).	97
6.1	Illustration of the different phrases of this dissertation	106
2	The figure above displays a violation related to access and disclosure of personal information	121
3	The figure above displays a violation related to unauthorized distribution of banking credentials	122
4	The figure above displays a violation related to online-to-offline threats	123
5	The figure above displays a violation related to misinformation	124
6	The figure above displays a violation related to the distribution of non-consensual explicit imagery	125

# Chapter 1

## Introduction

### 1.1 Problem Motivation

There are now almost 6.3 billion people across the globe using devices that are reliant on algorithmic and data-driven technologies (ADDTs) [55]. Access to ADDTs has become critical for accessing information, helping with decision-making, and connecting with others [8, 9, 34, 113, 183]. At the same time, the emergence of these technologies has altered the nature of safety by replicating and exacerbating existing patterns of injustice [153].

Algorithmic and data-driven technologies have now permeated multiple aspects of our lives making it difficult to be disentangled from the effects of its abuses as well. The potential risks to the human right to be free of physical, social, and physiological harms are high [139, 156]. Online spaces are being weaponized at exponential rates from multiple actors [86]. Individuals intimidate others with derogatory and demeaning language; unbeknownst to many, companies unfairly collect massive amounts of personal data and carry out extensive privacy abuses; state actors leverage online spaces to perpetrate dangerous misinformation and manipulative campaigns [169, 198, 158]. Being safe online is no longer restricted to the constraints of the technological system either as the threats spill over into our physical world too [150, 173].

Moreover, the opaqueness in these technologies come into questions as stakeholders massively under-serve some communities and continue to underestimate the growing number of bad actors who have learned to quickly game systems. In response, scholars and companies have attempted to address this problem by leveling the playing field with a focus on equality: all uses are afforded the



same resources and opportunities for risk mitigation [26, 10, 194]. However, this approach adopts a narrow socio-political perspective that misses the global diversity of the modalities in which harms may manifest. These digital technologies raise challenges not only to equality (being treated fairly), but also to equity (having the appropriate resources you need to achieve a fair outcome) [121]. As a result, internet users continue to face harms at exponential rates and vulnerable groups face harms at disproportionately higher rates [27, 136, 155]. Addressing issues of injustice is challenging and it could be further complicated when voices are excluded [43, 12]. Within this domain, the research, policies, and design of safety countermeasures, have been largely dominated by researchers from western, educated, industrialized, rich and democratic (WEIRD) nations with imbalances regarding the data sets used in models and systems as well as the representation of socio-cultural groups [174]. Costanza-Chock explains that "design mediates so much of our realities and has tremendous impact on our lives, yet very few of us participate in design processes" [42]. If we want to build better and safer, it is important to consider the increasing use of these tools and shed light where powerful actors misuse and abuse algorithmic technologies, violate human rights, and harm marginalized communities in all parts of the world.

I adopt an approach that echoes the concept of research justice [90]. Utilizing the research justice space allows for a more nuanced understanding of how fair and just outcomes could be designed and equitably distributed. This approach is centered on the idea of inclusion by centering the voices of those normally sidelined in design, both in the development of artifacts and understanding the processes that lead to artifact development. Thus, the employed approach emphasizes more just representation throughout design processes. In this dissertation, I lean on my lived experiences as a person from the Caribbean to uncover patterns of abuse in online spaces where Caribbean citizens reside and I work with local collaborators to discover how digital interventions could be designed to offer fair, just, and equitable solutions.

## 1.2 Research Objectives

There are multiple components that affect the development of safe digital interventions. To capture the complex interplay between different socio-ecological components, I use a multi-level framework based on ecology to understand the relationship between societal inequities and ADDTS [126, 32] (see Figure 1.1). The framework is bi-directional, i.e., the behaviors or experiences of an

individual can have ripple effects that extend to others within their periphery such as their social network or community, which in turn could affect the development of policy. Likewise, top-down effects stemming from policy (or the lack of policy) could influence community movements, the extent to which companies take action to issues, which affects the community and the experiences of individuals.

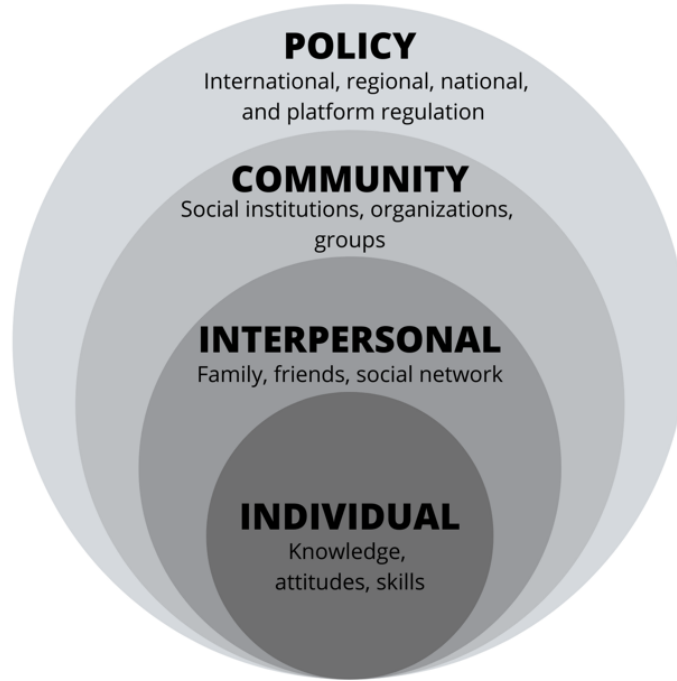


Figure 1.1: Socio-ecological model (adapted from [126, 32]) of the multi-level individual and societal factors that influence online safety.

In response to these challenges, this dissertation aims to answer the following questions:

- *How do Caribbean citizens perceive, evaluate and mitigate harms from ADDTs?* (**Chapter 3**)
  - **RQ1**: What is the prevalence of harms caused by or facilitated through ADDTs?
  - **RQ2**: Which harms are perceived to be the most concerning?
  - **RQ3**: How does users' coping strategies influence their adoption of protection behaviors?

- *What legal protections and opportunities for justice are available to Caribbean citizens?* (**Chapter 4**)
  - **RQ4:** Which countries have legal protections in place to support online safety?
  - **RQ5:** Across the region, which types of online threats are criminalized?
- *How could systems account for imbalances in opportunities for justice in ADDTs?* (**Chapter 5**)
  - **RQ6:** How do different justice-oriented countermeasures influence users' perceived safety within online communities?
  - **RQ7:** How does personalization affect users' perceptions of justice-oriented countermeasures?
  - **RQ8:** How do justice-oriented countermeasures influence the adoption of protective behaviors within online communities?

Throughout this dissertation, I present three key studies where I have worked on the ground with local organizations and collaborators to ensure the narrative is led by the community the research hopes to serve. In the first study, we conducted a survey with 551 Caribbean participants across 15 countries with the goal of understanding what motivates persons to adopt protective behaviors that address online threats. We show that by conceptually defining protective behaviors based on the threats that they address, it exposed how the perceptions of threats influence the adoption of online safety mechanisms while uncovering distinctions in perceptions depending on the type of harm. In the second study, we conducted a comparative study of the regulatory approaches to criminalizing violations of online safety. The study revealed that across the Caribbean there is a fractured approach to offering and implementing protections which in turn results in gaps in the enforcement of the laws. Based on the findings of these studies, I developed and evaluated the design of justice-oriented safety countermeasures.

## Chapter 2

# Related Work

The purpose of this chapter is to present core theoretical foundations relevant to the research undertaken within this dissertation. First, I reflect on threats that arise throughout the development cycle of ADDTs focusing on the input, output, and system-level interactions. I then reflect on theoretical approaches to respond to these threats and their influence on the design of digital safety mechanisms. Lastly, I expound on prior scholarship that integrates socio-technical perspectives, and I conclude by summarizing research gaps within this body of work.

### 2.1 Understanding Online Safety Threats

In the content of this dissertation, *safety* in digital spaces aligns with existing definitions within human rights frameworks and social computing research where safety is characterized by the absence or significant reduction of threats including emotional, physical, social, or psychological threats [155, 77]. There are ongoing investigations that adopt a more holistic perspective of safety outside of the silos of specific sub-domains such as harassment, cyber-stalking, privacy, or security. Instead, exploring multiple types of threats offers the opportunity to deeply understand the digital ecosystem to identify prevalent threats, stages in the development cycle with heinous abuses, and groups of persons who may be particularly vulnerable to safety risks both in digital and post-digital spaces [139, 150, 26, 156]. I extend this comprehensive lens to shed light on vulnerabilities to users' safety that could arise throughout the entire development cycle of ADDTs. This cycle is largely grouped into three main categories:

- *Input*: the phase where data is collected
- *Output*: the phase where data is processed and presented to users
- *System-level interactions*: the phase where ADDT components allow for interaction with end-users

In the subsections below, I discuss the three stages in the development cycle through an exploration of abusive behaviors that reduce the benefits of using ADDTs.

### 2.1.1 ADDT Threats: Input Level

Many people use ADDTs to connect with loved ones, stay updated with world news, and share personal updates, beliefs, and emotions [172]. The wealth of information collected in these systems makes it possible for companies to develop extensive user profiles [10, 19] and allow third-parties access to fine-grained information [98, 6]. While there are definite benefits to sharing personal information across ADDTs [114], there are also concerns as to whether these technologies collect more information about users than needed [201] and whether this information harvesting is ethical and transparent to users [58].

This data could potentially be used to make inferences about users’ behavior, socio-economic status, and even their political leanings [24]. Moreover, as more algorithmic and data-driven technologies are embedded into more day-to-day interactions, it becomes exponentially difficult to escape the wide reach of the pervasive nature of data collection that feed these systems. Social systems, such as social networking sites, are often used as a gateway to the Internet [161, 186]. Therefore, these proving safe experiences on these systems would be paramount. This position becomes of particular importance when the ADDT input could significantly affect users’ safety both digitally (in primarily online spaces facilitated by technology) and post-digitally (effects of using technology that spill over into the physical world). A prime example of this stems from sensitive data such as those collected from mobile fertility applications. These applications collect, process, and share information about users’ reproductive potential. Mehrnezhad and Almeida investigated the prevalence of leaks in fertility apps and that found that sensitive data is commonly ”mismanaged, misused, and misappropriated” [125]. Although data may appear to be localized to end users, (1) the extent of data collection may not be obvious nor are users offered the opportunity to have a granular level of awareness about active collection, and (2) data may be unknowingly shared with or

sold to partners and third parties including advertisers. Given variations in protections and rising concerns in the potential for legal prosecution regarding women’s rights, the protection of extremely sensitive personal data becomes paramount [180].

Compared to the web 1.0 era of the internet, technology companies now have access to richer data sources and more detailed personally identifiable information (PII) that can be used as attributes to make inferences about users [147]. Additionally, unbeknownst to many, data sharing among partners is extensive. For example, purchasing the data on 70 million US households enabled Facebook to tailor ads to specific audiences based on these users’ purchasing history on other sites [189]. While this has direct consequences for ADDT users’ online safety and well-being, users have a difficult time knowing who has this data and what is being done with it [2], since there are often no visual cues that indicate if or when this data is being shared.

In response to this problem, researchers have developed means to identify applications that leak the input attributes in ADDTs that are considered Personally Identifiable Information (PII) and identified methods to help reduce risk exposure (e.g., [5, 148, 75, 151, 171]). This effort contributes significantly to an academic understanding of leaks and violations in the ADDTs ecosystem.

In a previous study, we developed pervasive but unobtrusive visualizations that enhance users’ understanding of the real-time data-sharing practices of apps installed on their mobile devices [192]. We conducted a user study evaluating different design prototypes of these visualizations. Figure 2.1 illustrates the different designs presented. The goal of the study was to increase users’ awareness of how their input (either explicit or implicit) was being used by systems in an effort to enhance accountability in ADDTs. By manipulating the structure and granularity of the information, we gained insights into users’ preference regarding these aspects, as well as the effect these aspects have on participants’ understanding of the visualizations and the information they contain. We found that our participants’ preference of information structure depended on their perceptions of privacy boundaries as characterized by Petronio’s theory of Communication Privacy Management (CPM) [142]; participants who considered apps to be appropriate co-owners of their personal information preferred the app-centric designs, whereas those who were more focused on the information being shared (regardless of the app) preferred the data-centric designs. These findings provide support for the design of tools that account for user difference, thus departing from the one-size-fits-all approach to design.

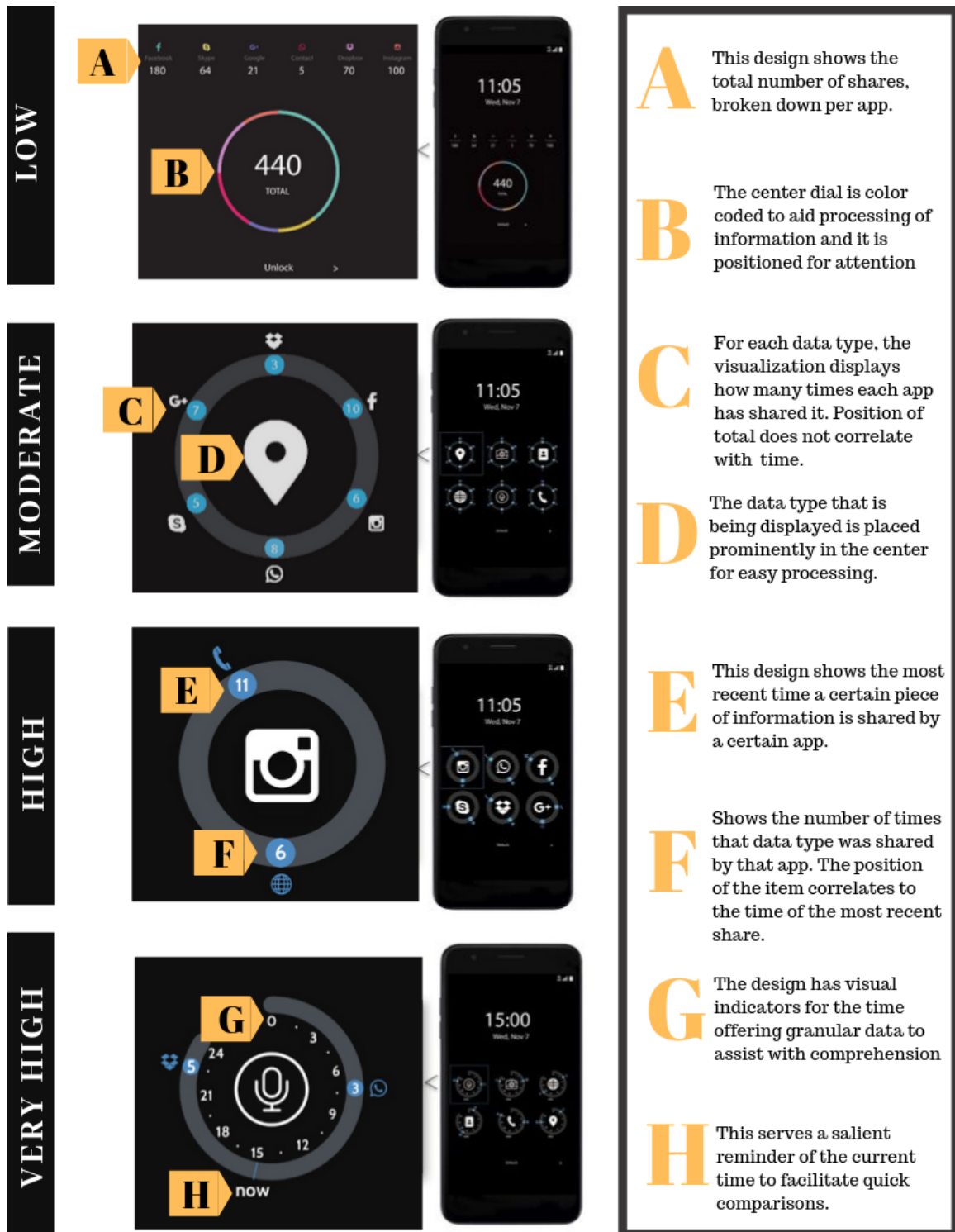


Figure 2.1: Annotated Designs: Designs varied in their level of granularity (low, moderate, high, very high) and the presentation style (app-centric versus data centric) with a total of eight designs. For each design shown, there is an identical design with the same level of granularity but different presentation style. From top to bottom: "Low granularity, app-centric presentation", "Moderate granularity, data-centric presentation", "High granularity, app-centric presentation", "Very High granularity, data-centric presentation".

### 2.1.2 ADDT Threats: Output Level

Once data has been harvested and processed, the output is presented to end users. However, users may struggle to cope with incorrect inferences or misunderstanding the algorithmic system. Recently, there has been a surge in interpretability and explainability research in Machine Learning and similar domains, acknowledging the importance and benefits of more interpretable systems [190]. Moreover, some studies have viewed interpretability in terms of providing explanations around the input parameters that most impact the output. Other works have explored incorporating transparency at the early stages of design with the hope that it would result in more interpretable systems [190]. Explanations allow users to better understand and interpret the rationale that leads to the output from ADDTs. Prior work has shown this can lead to improved trust, transparency and user engagement [165, 57, 124, 144, 179, 76]. For example, Kouki et al. [101] present a hybrid recommender system that is built on a probabilistic programming language, and they demonstrate that explanations improve the user experience of the recommender system. Likewise, Friedrich et al. describe a taxonomy of explanation approaches, taking into account different dimensions like the style (e.g., collaborative, knowledge, utility or social explanation style), paradigm (e.g. content-based, knowledge or collaborative based) and the type of preference model [62]. In [176], authors create explanations through capturing the interactions between users and their favorite features by constructing a feature profile for the users. Moreover, they use a feature-weighting scheme to reveal those features which better describe a user and those which better distinguish that user from the others. In addition, different visualization techniques are proposed for providing explanations for the generated recommendations, such as interfaces with concentric circles [92, 134], and pathways between columns [29]. In a similar study, Dominguez et al. experiment with different interfaces with different levels of explainability and different algorithms for artistic image recommendation [52].

In our prior work, we considered the justification style in terms of the type of user *question* it answers (see [191]). Specifically, we studied the effect of providing justifications in intelligent systems by employing three different justification styles, as well as interactions with two different recommendation algorithms: one with high accuracy, and another with lower. Our results showed that *why* justifications (rather than *why not*) significantly influence users' perception of system transparency, influences perceived control, and in turn affects users' trusting beliefs and intentions. Beyond recommender systems, explanations and justifications have been studied for broader ADDTs,



most notably in the field of explainable artificial intelligence (xAI) [70, 127, 69] as well as expert and systems[146, 164], adaptive agents [66], and context-aware technologies [112].

Prior work has explored methods to detect *why* users were being recommended particular output [105, 48, 83, 16]. Andreou et al. investigated the effectiveness of ad explanations on Facebook [10]. They found that explanations on the platform were often incomplete, misleading or vague. Similarly, Eslami et al. found users preferred interpretable “non-creepy” explanations for ads on social media [56]. However, balancing *how* to provide explanations that align with what consumers want is a challenge, as revealing too much or too little about the algorithmic process has been shown to both negatively impact system trust [96].

Regardless of the approach, the underlying goal of this field of research is to provide end-users with sufficient information to assist them in identifying misbehavior and understanding why an ADDT produced an output. Figure 2.2, presents a classification scheme for explaining output [195].

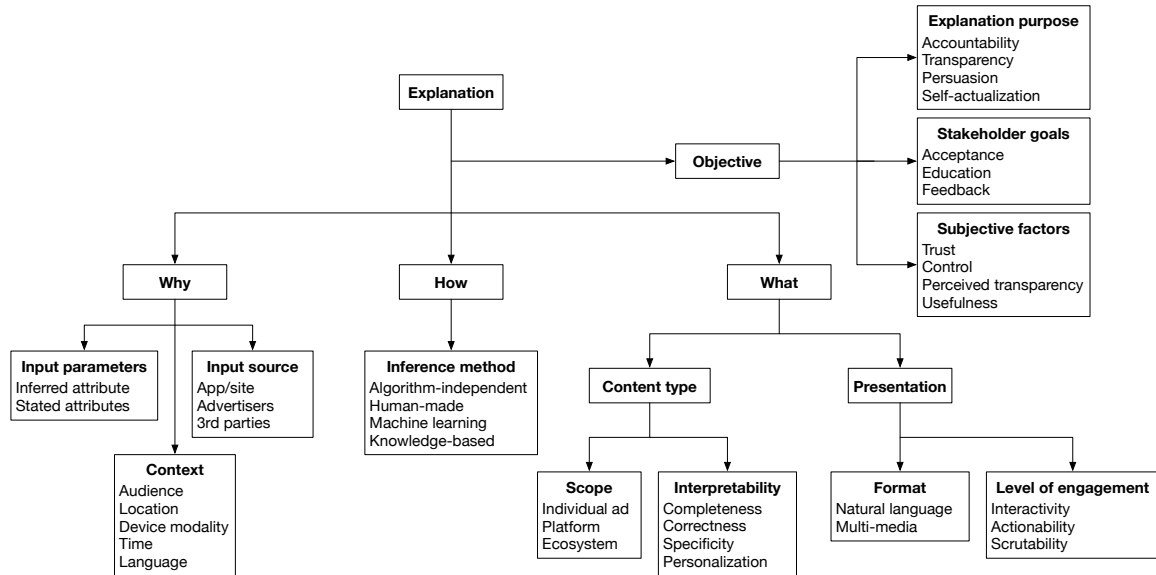


Figure 2.2: Explanation classification scheme

### 2.1.3 The Impact of System-level Interactions

The negative impact associated with the input and output components in the ADDT development cycle could be further compounded by ill-indented actors. ADDTs have been leveraged to

spread hate speech, the non-consensual distribution of explicit photos, stalking, fraud, and a host of harassing behaviors [26, 107]. While these have historically been depicted as an outlier or fringe behavior, abusability and abusive behavior have been endemic in online space since their inception [173]. Moreover, often those who are already vulnerable are disproportionately affected by abusive behavior in online spaces [203, 61, 82]. In response, HCI scholars have made considerable strides towards understanding system-level threats [27, 155]. However, advocates have still raised concerns about where the responsibility lies in the design of tools to curb these types of behaviors. Technology companies have typically adopted a “neutral” approach to governance that effectively absolves them of the responsibility to adjudicate harm [63]. Most platforms maintain community guidelines which categorizes the types of behaviors that would be deemed unacceptable [87]. Through content moderation, companies rely on a bevy of AI-supported and human methods to determine violations [64]. The scale and variety of harms have motivated technology companies to harvest the power of artificial intelligence (AI) approaches to prevent, detect, and rectify harms. Some advocates have propped AI as a panacea that would identify misinformation, hate speech, pornographic material, and other platform violations in a quick and fair manner before the offending content is uploaded for others to see. This idea was unofficially tested on a large-scale during the coronavirus pandemic as many companies relied on automated content moderation as their human moderators were sent home. It was a failure. Human rights journalists who rely on social media to document injustices, saw multiple accounts of activists being shut down without the option to appeal the decision [160]. Meanwhile, problematic content remained untouched as human moderators were not able to serve as arbitrators and determine the nuances that would indicate platform violation. Consequently, the numbers for the removal of high risk content, like child exploitation and self-harm on Facebook, were 40% lower in the second quarter of 2020 [160]. The results of this test raise questions about the reliance on solely automated approaches. Is it fair to rely on technology to decode complex human issues that humans have difficulty with - especially when the matters include systematic oppression, race relations, political power-plays, and economic dynamics, etc.? Over time, public pressure has steadily increased, calling for companies to take more action against harms perpetrated and facilitated by their products [72]. In the United States, debates have sparked around Section 230 of the Communications Decency Act, which protects tech giants like Meta and Twitter from prosecution based on harmful content distributed on their platforms [72]. Even when regulatory protections are in place, regulations should not be used to perpetuate further harm. Saki and Sambuli describe

how, in Uganda and Brazil, respectively, anti-pornography laws and defamation lawsuits have been used to punish women for being online rather than protecting them [163]. These are critical considerations for keeping gender and sexual minorities (GSM) in the region safe. Safeguards should be implemented to prevent any particular governing or political body to unfairly use frameworks to target groups of people. Thus, there are still challenges remaining around who should be responsible misbehavior and how institutions should enhance their offerings of protective solutions.

## 2.2 Theoretical Considerations

The ubiquitous nature of online threats and its associated growing concerns to user safety have motivated multiple stakeholders to consider multi-disciplinary approaches to address threats. Social media companies have widely adopted frameworks centered around content moderation where more punitive countermeasures are applied to remove content or ban users based on violations of established guidelines [87]. Meanwhile, advocates and regulators continue to apply increased pressure for social media companies to accept more accountability through policy that would strip these companies of immunity from prosecution over most of the content users publish on their platforms [72, 18]. In the midst of these discussions, scholars have leaned towards theories of justice to inform the design of safety countermeasures that would best serve the needs of users. In the following section, I provide an overview of relevant scholarship that offers theoretical foundations that aim at understanding justice and its application to the design and development of countermeasures. I also explain how these theories influence the work conducted in this dissertation.

### 2.2.1 Theories of Justice in HCI

This work aligns with the core principles of *research justice* by uplifting the voices of communities and scholars who have historically been marginalized in the definition and production of knowledge [90]. Research justice posits that people who are unable to claim their experiences and begin to internalize dominant narratives feel disempowered to challenge power [90]. It emphasizes strategies for the transformation of policy and encourages underrepresented forms of knowledge development while centering community members as the experts. The underlying principles highlighted in by this theory closely aligns with social science but also connects with recent trends in HCI scholarship to dismantle power imbalances in dominant research narratives. For instance, re-

cent efforts within the HCI community challenge the focus on Anglo- and Euro-centered narratives [1, 133, 86].

HCI literature offers a variety of approaches through which both justice and design are considered. Sasha Costanza-Chock and the Design Justice Network have led considerable efforts in highlighting the role design plays in uncovering exploitative, abusive, and oppressive systems [42, 43]. Similar to the principles of research justice, *design justice* places marginalized scholars and their communities at the forefront of scholarly and design efforts. In this approach, design is not limited to the development of artifacts but also the design processes. As such, community-led and participatory efforts are essential in this approach.

Within the context of online safety, the core implication of these two justice approaches would be acknowledging the power held by populations who disproportionately experience harms, thus, gaining authority in the identification of injustice given their lived experiences. This calls for a deeper understanding of these concepts through knowledge sharing from those who have been disproportionately harmed. Scholars like Abeba Birhane argue that the "starting point toward efforts such as ethical practice in machine-learning systems or theories of ethics, fairness, or discrimination needs to center the material condition and the concrete consequences an algorithmic tool is likely to bring" [25].

Scholars have also adopted theories that address systemic challenges that people face due to aspects of their identities or legacies of structural imbalances. For example, *Post-colonial HCI* is centered on addressing the challenges associated with the history of colonialism by dissecting unjust power dynamics in the development of technology [84]. In this work, Irani et al. recognize shifts in emerging technologies but acknowledge "colonial relationships may have dissolved, and yet the history of global dynamics of power, wealth, economic strength, and political influence shape contemporary cultural encounters" [84]. This lens is relevant to the work in this dissertation, as lasting effects from decades of colonialism has shaped the development of countries in the region. Beyond cultural or historical lens, theoretical underpinnings from *feminist HCI* have shed light on issues of reproductive justice, sex work, sexuality and social change by extension [93, 15]. More recently, Mariam Asad introduced *prefigurative design* as a framework focused on justice and equity through community-based collaborations [12]. This work is heavily influenced by criminal justice models such as transformative justice which emphasize the development of counter-institutions.

## 2.3 Socio-technical Perspectives

Although companies have opted to begin deploying more safety countermeasures, the design of these measures have often been grounded in predominantly universalist design principles which falls short in addressing the varied and intersectional needs of a globally diverse population [158, 157, 42]. Scholars have long advocated for the acknowledgement of individual differences that influence variance in behavior and safety needs based on characteristics such as nationality, age groups, and genders [86, 110, 182, 155]. As such, social computing researchers have begun highlighting the importance of "de-centralizing" largely Anglo- and Euro-centered narratives [1, 133, 86] narratives that may dominate design spaces. Instead, a recent push in research direction has been focused departing from Western, Educated, Industrialized, Rich, and Democratic (WEIRD) cultures [174, 115] by amplifying the voices of non-WEIRD researchers and including more variation in the populations considered in research.

Although there is extensive literature available on online threats or harms, there are very few studies that focus on users' protective behaviors in non-WEIRD countries and even fewer works that particularly focus on populations like the Caribbean [174]. However, empirical findings point towards the importance of diversifying the populations considered in research to more effectively uncover and address harm at a larger scale. Jiang et al. investigated the perceptions of harm across eight countries and found unique country differences in perceptions related to severity across multiple harms [86]. The empirical findings confirm significant country-to-country variations in what was perceived to be severe, yet, the design of countermeasure continue to adopt a one-size-fits-all approach although data points to the need for new methods to prioritize and customize safety experiences [86]. Thus, researchers have pointed to alternative methods for justice that could be responsive to user needs while also adopting a socio-technical and culturally-respectful lens. Aligning with this view, Im et al. investigated gendered differences in the perceptions of online harm across 14 geographic regions that have traditionally been understudied such as Mongolia, Cameroon, and the Caribbean [81]. The study revealed regional preferences regarding restorative approaches to justice such as payment versus more punitive approach such as banning or content removal. Beyond understanding attitudes and preferences towards safety-related countermeasures, researchers have also explored the influence of social-technical factors such as culture-related indicators (for example country of residence, language, cultural dimensions) to predict decisions that would help maintain

safety online [109]. Li et al. found cultural indicators significantly improved prediction accuracy related to the acceptability of personal data disclosure [109].

These works contribute to a better understanding of the way socio-technical indicators are embedded into technologies and deeply intertwined across society at-large. However, they also limit their scope of examining experiences of very specific harms (such as harassment) or investigating these phenomena with very specific populations.

## 2.4 Research Gaps

Overall, there has been considerable strides towards the understanding of user safety needs regarding ADDT usage. However, the above-mentioned work also highlights multiple limitations. At the forefront, there are very limited works that explore online safety perceptions and protective behaviors in the Caribbean. The work completed in this dissertation is the first to comprehensively investigate this topic within the region by focusing on key limitations in the literature on online safety.

First, the investigations of harms are largely studied in silos rather than a comprehensive approach that would assist in building theory about experiencing and responding to harms. This limitation in the literature is addressed in Chapter 3 by operationalizing a holistic view of online safety. Secondly, the scope of the response is often limited to technical capabilities with minimal regard for societal influences. Response to harms that occur in the digital space are influenced by multi-level factors such as interpersonal relations, culture and the community, and available regulatory policy (as illustrated in 1.1. Thus, Chapter 4 examines non-technical forms of responses such as regulatory protections, discusses the state of online safety policy in the region and how this in turn affects the design and development of technical safety countermeasures.

Lastly, existing scholarship on the application of justice theories in the design of countermeasures have largely been theoretical. Therefore, this work builds on foundational literature on justice theories by applying those principles with empirical work done in chapters 3 and 4 to produce and evaluate informed design artefacts in chapter 5.

## Chapter 3

# Exploring Safety Perceptions and the Prevalence of Threats

*This paper was accepted to the 2022 ACM CHI Conference on Human Factors in Computing Systems, under the title "Many Islands, Many Problems: An Empirical Examination of Online Safety Behaviors in the Caribbean" [193].*

### 3.1 Overview

To understand the prevalence of online threats in the Caribbean, we conducted a large-scale survey throughout the region to develop a deeper understanding of how people in the region perceive, evaluate, and mitigate threats to their online safety.

Research across multiple disciplines has shed light on the incredibly varied and widespread nature of digital harms. Social media platforms have served as easily accessible mediums for people to celebrate major life milestones, maintain interpersonal relationships, engage in discourse, and be an outlet for coping with crises and grief [8, 9, 34, 113, 183]. At the same time, these platforms have been central to the proliferation of harmful behaviors online. Individuals target others with inflammatory language or insults; unbeknownst to many, companies unfairly collect massive amounts of personal data and carry out extensive privacy abuses; state actors leverage the online space to perpetrate dangerous misinformation and manipulative campaigns. Within the context of social media, risks

to online safety refer to a broad spectrum of threats relative to security, privacy, harassment, and well-being that are typically studied in silos and focused on online interactions. However, those lines are blurred in real-world experiences, where social media users are often faced with the challenge of navigating risks online and trying to avoid spill-over effects into their physical worlds. With this notion in mind, researchers have argued that the concept of “safety” in digital spaces should be seen as protection from harm (i.e. perceived threats, injury, or unwanted outcomes) [156].

Moreover, the perception of dominant (often Western) frameworks as the standard for the implementation of safety mechanisms fails to account for imbalances, inequalities, and injustices in non-Western civilizations like the Caribbean. Thus, in this survey study (N=511), we investigate the extent of online safety threats throughout the Caribbean region, current protective behaviors being employed, and differences in users’ perceptions of various types of harms. Given the complexity of what it means to be safe online, we examine a wide range of harms related to security, access and disclosure, harassment, and online-to-offline threats. We propose a conceptual framework based on the Protection Motivation Theory (PMT) [122], to understand what factors motivate Caribbean social media users’ safety intentions. To explore the relations between these factors, the paper addresses the following research questions:

**RQ1:** Which types of threats are prevalent?

**RQ2:** Which threats are perceived to be the most concerning?

**RQ3:** What role does users’ threat and coping appraisals play in their intention to adopt protective behaviors?

## 3.2 Background

We draw on prior research in two main areas: harmful online experiences and protective behaviors. Specifically, we focus on experiences that define safety, and factors that influence the adoption of harm mitigation strategies. Additionally, we describe cross-cultural considerations within this context. We then offer insights into the theoretical foundations our work is centered around. Last, we present our hypotheses for the study.



### 3.2.1 Online Threats and Safety Protection

#### 3.2.1.1 Adopting a Wider Lens on Online Safety

Inherently, the design of social network systems encourages online interactions, which has proven to have immense benefits to discourse, social support, and overall well-being [8, 9, 34, 113, 183]. It should be noted, that these types of interactions also create severe vulnerabilities for users. Threats to our safety online could result in injury, loss, harm, or deprivation. Prior work examining perspectives on safety often focus on either technical and/or relational views. Technical perspectives are focused on concerns about system vulnerability and information flows. For example, phishing scams, virus protection, security practices, and concerns about access to personal information. Relational safety concerns are centered around interpersonal harm, such as bullying, hate speech, and harassment [27, 155]. Unfortunately, alarming trends in the rates of threatening online content point to a growing number of malicious actors who have learned to weaponize systems for threatening activities [86]. These evolving threats and vulnerabilities require an expansion of our understanding of online threats and what protections we should consider. For example, the harassment of women on digital platforms has ballooned to such a heightened threat that experts at the United Nations argue it is now a human rights violation [136]. In a similar light, misinformation online has been shown to influence elections and highlighted its potential as a viable threat to democracy [198].

In response, HCI scholars have made considerable strides towards understanding online threats, and many researchers now acknowledge the complexities of what it means to be safe online. Rather than investigating very specific elements of safety threats in isolation, Redmiles et al. argued that adopting a wider lens allows us to see the entangled nature of day-to-day experiences that influence users' perceptions of safety [150]. Researchers have gradually moved beyond examining solitary harms and instead exploring dimensions of online harms in an effort to understand possible approaches to harm mitigation. In this light, Scheuerman et al. presented a framework that focused on four types of harm—physical, emotional, relation, and financial [156]. The work highlights the importance of investigating multiple harms to better understand how they relate to each other. In our study, we define safety along the lines of Pater et al. [139], referring to freedom from emotional, physical, and social harm that may be caused by—but is not always caused by—abusive behavior.

Although behavior on social media is reflective of societal behaviors, these platforms have been used to facilitate and amplify threats. As such, scholars have called for an in-depth review

and redesign of socio-technical systems that departs from the approach to development focused on building fast and fixing later [173]. Soltani argues that building safer technology requires a comprehensive testing of platforms' vulnerability to being abused and that teams need to adopt abusability testing [169]. To provide a more holistic view of the threats affecting social media users, significant strides must be made to investigate wider descriptive characteristics of those who experience vulnerabilities. Extant research has shown that people from different countries, age groups, and genders behave differently online [86, 110, 182]. However, much of the work that focuses on protective behaviors has (1) largely been focused on Western, Educated, Industrialized, Rich, and Democratic (WEIRD) cultures [174, 115], and (2) focused on elements of safety rather than perceptions that motivate safety. In contrast, this work builds on recent efforts within the HCI community that challenge the focus on Anglo- and Euro-centered narratives [1, 133, 86]. Although there is extensive literature available on online threats or harms, very few studies that focus on users' protective behaviors in non-WEIRD countries and especially the Caribbean [174]. Recently, Jiang et al. investigated the perceptions of harm across eight countries and found unique country differences in perceptions related to severity across multiple harms. Our work complements and expands on this work by considering 15 countries across a region often excluded in HCI research.

### **3.2.1.2 Online Safety in the Caribbean**

The Caribbean is a group of heterogeneous countries. Historical connections forged by colonialism have created a region that prides itself as a melting pot with diverse backgrounds in political stature, culture, and economic development. Although the region is strongly tied by culture, wide variations exist, and each country has unique attributes and challenges even though they are geographically closely located. These differences could be illustrated in dual-governed islands such as St. Martin/St. Maarten. On the 37 square miles island, the north is controlled by the French while the south is Dutch. There are no physical borders but both sides practice different laws, have different languages, and adhere to different cultural practices. On another scale, Caribbean countries often work collaboratively through organizations such as the Caribbean Community (CARICOM) in order to have a more unified voice. Thus, the region may operate collectively on international matters similar to the European Union but still maintain very granular differences due to socio-economic and historical factors. Despite these differences, regional leaders have been vocal about the need to adopt more technology-driven economies to maintain global competitiveness and promote

sustainable social development. As the region’s economies continue to face disruption to traditional industries such as agriculture and tourism, it is critical to take a proactive rather than reactive approach to aid the transition to more digital societies. This transition to more digital societies may bring its own problems, though, such as an elevated threat to users’ online safety.

Undoubtedly, online safety and safety-focused movements are gaining momentum globally [150, 158, 86] including within the Caribbean region [33]. Calls in this domain have largely been driven by regional leaders who have collectively acknowledged the transition to more digital societies could create new vulnerabilities that need to be considered earlier rather than later [36]. Caribbean leaders pushed for the creation of the Caribbean Community Implementation Agency for Crime and Security (CARICOM IMPACS)<sup>1</sup> which leads multiple initiatives that have resulted in wide-reaching discussions and training that improve capacity building related to enhancing the detection and investigation of violations in the digital space. Yet, there is a lot to be done before governments in the region can offer a united approach to protection in the digital space. From a legislative standpoint, protections are inconsistent and as of the end of 2021 only 10 countries in the region have enacted substantive data protection legislative policies [135]. The goal of CARICOM, is to utilize the collective power of its member states throughout the region to promote consistency and shared benefits. And although their goal is to implement a GDPR-style approach to offering regulatory protections, privacy experts assessing the region’s response to online threats have concluded that the “Caricom is where the EU was at in 1988 in developing GDPR” [119].

Beyond, governmental efforts, very few research has been conducted on online safety in the Caribbean. The few studies that have covered this region are limited to very specific threats or focused on one country. For example, Thakur investigated how technology was being used to further facilitate gender-based violence in Jamaica [178]. The study found that 65% of respondents witnessed abuses online and 71% thought it was a major problem. Similarly, Smith and Stamatakis explored factors that affect cyber-crime victimization for cyber-bulling and unauthorized access in Trinidad and Tobago [167]. Both studies focused on the occurrence of very specific harms happening in one country in the region and did not explore protective behaviors. In this study, we attempt to fill this gap by investigating factors affecting safety behaviors of Caribbean citizens across the region. To do this we employ Protection Motivation Theory.

---

<sup>1</sup>CARICOM IMPACS: <https://caricomimpacs.org/cyber-security/>

### 3.2.2 Protection Motivation Theory

Protection Motivation Theory (PMT) [122, 152] provides a critical lens to examine how and why people decide to engage in protective behaviors in potentially threatening situations. The theory proposes that behavior is influenced by users' appraisal of threat and their coping appraisals regarding this threat. Threat appraisals are conducted to determine an individual's overall perception of danger, and are determined by the perceived severity and perceived vulnerability associated with unsafe situations or behaviors. Similarly, coping appraisals are conducted to determine an individual's ability to respond to the threat, and are determined by the response efficacy and self-efficacy associated with carrying out safe behaviors. Both the threat and coping appraisals are mutually inclusive. Both types of appraisal must occur for individuals to eventually perform the protective behavior: If a threat is not perceived to be severe, unlikely to occur, or if users felt like nothing could be done about the threat, no protective motivation would emerge and ultimately there would be no change in behavioral intention.

Within the context of social media, safety mechanisms are often available to assist users in the event of specific threats. However, it is ultimately up to the user to determine whether or not those mechanisms will help them feel safe while interacting online. Therefore, an individual's assessment of their disclosure patterns on social media may be influenced by an assessment of the benefits and threats of engaging when it is potentially unsafe. The objective of the current study is to investigate factors contributing to information disclosure when users feel safe or unsafe. As illustrated in Figure 3.1, when all of the four appraisal components are put together, they are deemed to influence users' level of safety protection. The model posits that the components have a linear relationship with protection motivation. Namely, as any of the variables increase, a higher level of protection motivation will occur. Thus, all of the individual variables are considered to be equally essential, rather than any one being of more importance than the others [152, 122].

Recent studies that have applied PMT in the context of online safety have investigated the motivation behind using computer virus protection [104], online privacy [204], harassment [118], predicting internet scam victimization [38] and digital security [162]. Therefore, it would be appropriate to apply PMT in examining how social media users manage risks related to their safety by adopting online protection behaviors. Unlike previous works, this study applies PMT to empirically measure a multitude of behaviors that contribute to safety, rather than focusing on one particular protective

behavior. In doing so, we demonstrate the importance of explaining safety practices as a whole and within the context of varying types of harms, as opposed to addressing but one particular context.

### 3.2.3 Hypothesis Development

According to PMT, threat appraisal, which is comprised of perceived severity and perceived vulnerability, acts as a determinant of whether one adopts coping responses [60]. A novel contribution of our study is the examination of prior experiences with safety threats and its association with such threats appraisals. Prior work has found that prior experiences serve as significant predictors in making decisions about online harms [39]. This likely happens because those who have personally been victims of safety harms are likely to understand the severe consequences associated with that threat [23].

For example, Mohamed and Ahmad found that persons who were victims of internet scams tended to build more knowledge about related severity and vulnerability [128]. Thus, we hypothesize that prior experiences with safety risks will influence users' perceptions of how much they can trust social media, while also affecting their awareness of the consequences of risk exposure, thus impacting their perception of the severity of that harm and their perceived vulnerability to it.

**H1:** Threat experience will have an effect on perceived vulnerability

**H2:** Threat experience will have an effect on perceived severity

**H3:** Threat experience will have an effect on perceptions of trust in social media platforms.

According to PMT, coping appraisals are formed from response efficacy beliefs (i.e. the belief that blocking a person on social media would protect them from additional harassment) and self-efficacy beliefs, which is the extent to which one believes they have the ability to successfully use to a safety tool (e.g. the belief that one could effectively use two factor authentication) [60]. This aligns with prior research which showed that the more people thought a harm was severe, the more likely they were to adopt positive attitudes towards protective behaviors [152]. Woon found that increased levels of perceived severity positively affected participants' security behavior [197]. Likewise, Johnston and Warkentin showed that the more people felt they were vulnerable to a threat, the more likely they were to consider the capabilities of protective mechanisms [89]. With this in mind, we present the following hypothesis:

**H4:** Threat experience will have an effect users' coping appraisal

As an individual experiences stronger attitudes towards how well a particular safety mechanism works in maintaining safety, they will be more motivated to engage in that protective behavior [118]. In a similar way, a user who is more confident in their ability to effectively use a tool is more likely to be positively motivated to engage with that tool [122]. Hence, we propose:

**H5:** Users' coping appraisal is positively associated with behavioral intention

**H6:** Perceived severity is positively associated with behavioral intention

**H7:** Perceived vulnerability is positively associated with behavioral intention

In human interaction, trust has been viewed as a critical factor in interactions involving risk, and the effect of trust has also been studied extensively in technological contexts [123]. Studies have shown that social media users are more likely to trust platforms that could keep them protected from safety harms [3]. Kim et al. illustrated how usable privacy policies predicted consumers' trust of a website [94]. Conversely, social media companies have faced increasing public pressure because of risks to users' safety, such as unfair data collection [200], harassment [187, 26], and overall concerns for better safety tools [150]. Based on these findings, we hypothesize the following:

**H8:** Trust in social network platforms is positively associated with behavioral intention

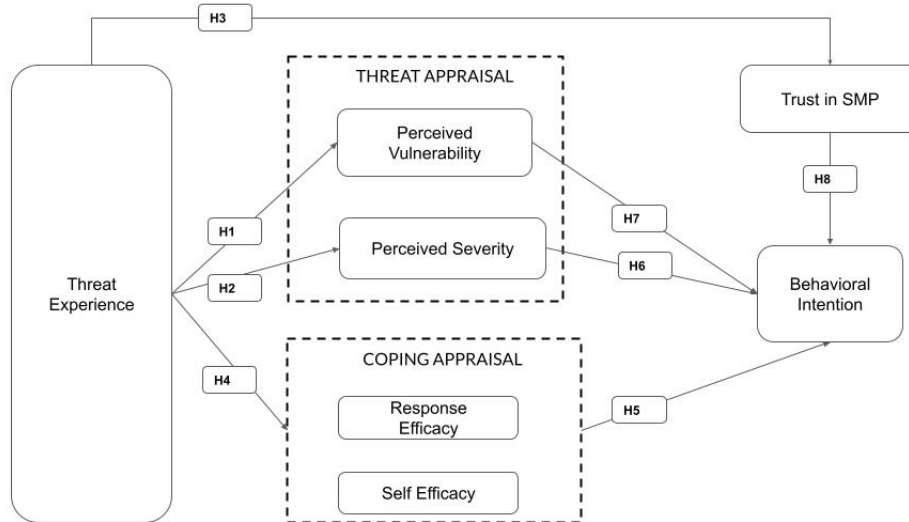


Figure 3.1: The figure above illustrates the proposed conceptual model for the study

### 3.3 Method

To test our hypotheses we conducted an online survey with 563 participants throughout the Caribbean region between March to June 2021. This study was reviewed as Exempt by our university’s Institutional Review Board. In the following section, we describe the methodologies adopted, the study procedures, and the recruited sample of study participants.

#### 3.3.1 Recruitment

Participants were recruited from a total of 15 English speaking countries in the Caribbean region: Anguilla, Antigua and Barbuda, Barbados, Bonaire, Cuba, Curaçao, Dominica, Grenada, Guadeloupe, Jamaica, Martinique, Saint Kitts and Nevis, Saint Lucia, Saint Martin, Saint Vincent and the Grenadines, and Trinidad and Tobago. The description of the demographics is included in Table 3.1.

Country	Male		Female		Non-binary		Self-describe		Prefer not to say		Total
	%	count	%	count	%	count	%	count	%	count	
Jamaica	24.14%	35	65.52%	95	0.69%	1	0.69%	1	8.97%	13	145
Saint Kitts & Nevis	27.27%	27	53.54%	53	0.00%	0	2.02%	2	17.17%	17	99
Dominica	14.29%	10	72.86%	51	0.00%	0	2.86%	2	10.00%	7	70
Barbados	28.13%	18	59.38%	38	0.00%	0	1.56%	1	10.94%	7	64
Saint Lucia	25.42%	15	61.02%	36	0.00%	0	0.00%	0	13.56%	8	59
Antigua and Barbuda	32.35%	11	47.06%	16	0.00%	0	2.94%	1	17.65%	6	34
Trinidad & Tobago	15.63%	5	53.13%	17	18.75%	6	0.00%	0	12.50%	4	32
Saint Vincent	37.50%	9	41.67%	10	0.00%	0	8.33%	2	12.50%	3	24
Grenada	20.83%	5	70.83%	17	4.17%	1	0.00%	0	4.17%	1	24
US Virgin Islands	66.67%	2	33.33%	1	0.00%	0	0.00%	0	0.00%	0	3
Saint Martin	50.00%	1	50.00%	1	0.00%	0	0.00%	0	0.00%	0	2
Anguilla	50.00%	1	50.00%	1	0.00%	0	0.00%	0	0.00%	0	2
Guadeloupe	0.00%	0	0.00%	0	0.00%	0	0.00%	0	100.00%	1	1
Martinique	100.00%	1	0.00%	0	0.00%	0	0.00%	0	0.00%	0	1
Bonaire	100.00%	1	0.00%	0	0.00%	0	0.00%	0	0.00%	0	1
Cuba	100.00%	1	0.00%	0	0.00%	0	0.00%	0	0.00%	0	1
Curaçao	100.00%	1	0.00%	0	0.00%	0	0.00%	0	0.00%	0	1

Table 3.1: Gender distribution per country. Countries in the second segment of the table were excluded from the analysis due to a low number of participants.

We recruited respondents by using a combination of online recruitment on social media, snowball sampling, and word-of-mouth techniques. We contacted community organizations within the region and posted in Facebook groups of the respective countries. The recruitment message requested participants who were currently residing in the Caribbean and used the Internet. Participants were required to be 18 years or older. On average, it took 19 minutes to complete the study. Respondents were offered \$5 USD in mobile credit to thank them for their time. The amount and

type of incentive was decided after conferring with local collaborators and speaking with persons during the pilot phase. All of the responses were anonymized and extra steps were taken to prevent re-identification. An attention check question was included to help to identify poor quality responses. In total, five responses were excluded from the analysis due to low quality, which left a total sample size of 551.

Mobile Application	Never Used it		Don't use it anymore		Haven't used it in a while		I'm using it now	
	%	count	%	count	%	count	%	count
WhatsApp	0.23%	5	1.50%	9	0.85%	8	17.72%	548
YouTube	0.18%	4	1.83%	11	3.72%	35	16.81%	520
Facebook	0.91%	20	8.65%	52	6.91%	65	14.00%	433
Instagram	2.67%	59	5.49%	33	7.77%	73	13.09%	405
Snapchat	6.53%	144	10.98%	66	12.34%	116	7.89%	244
Tik Tok	10.24%	226	8.15%	49	7.55%	71	7.24%	224
WhatsApp mod*	11.60%	256	9.98%	60	5.11%	48	6.66%	206
Pinterest	7.48%	165	8.65%	52	17.87%	168	5.98%	185
Twitter	8.57%	189	14.81%	89	13.51%	127	5.33%	165
LinkedIn	13.24%	292	11.48%	69	10.85%	102	3.46%	107
Reddit	19.63%	433	5.99%	36	6.81%	64	1.20%	37
Tumblr	18.72%	413	12.48%	75	6.70%	63	0.61%	19

Table 3.2: Description of the frequency of app usage among all participants. Note that "WhatsApp Mod" represents WhatsApp FM, GB WhatsApp or any modified version of WhatsApp\*.

## 3.4 Findings

We organize our results by the initial research questions outlined in section 3.2.3 and present the findings related to our proposed conceptual model.

### 3.4.1 RQ1: What threats are prevalent throughout the region?

We first explored how participants across the region experienced threats to their safety. Overall, 92% of respondents reported having experienced a threat to their online safety on at least one occasion. When comparing the prevalence of the different types of threats, risks regarding access to personal information and disclosure were the highest, with 43% of participants reportedly having experienced this type of threat. Additionally, 30% of the respondents reported having experienced security related threats, 35% experienced harassment-related threats, and 32% reported threats that transferred from the online to the offline space. In Figure 3.2, we show the overall distribution of threats among all participants. This visualization is revealing in several ways. The top three experienced threats were spread across different groups of threats, rather than belonging any one



type of threat. The most prevalent threat was related to targeted advertising as 58% of participants reported having experienced their personal information being collected and used to send unwanted ads on social media. The second highest occurrence was being sent unsolicited explicit content (55% of participants reported having experienced this harm). We also observe high instances of prior experiences with potentially compromised login information (54.45% of participants reported having experienced this harm).

We subsequently adopted a more focused observation to distinguish between differences in victimization rates across the region. Figure 3.3 shows a general trend of similar victimization rates among all threats. However, we note a trend of consistently higher reported experiences among participants from St. Vincent and consistently lower rates among participants from Trinidad and Tobago.

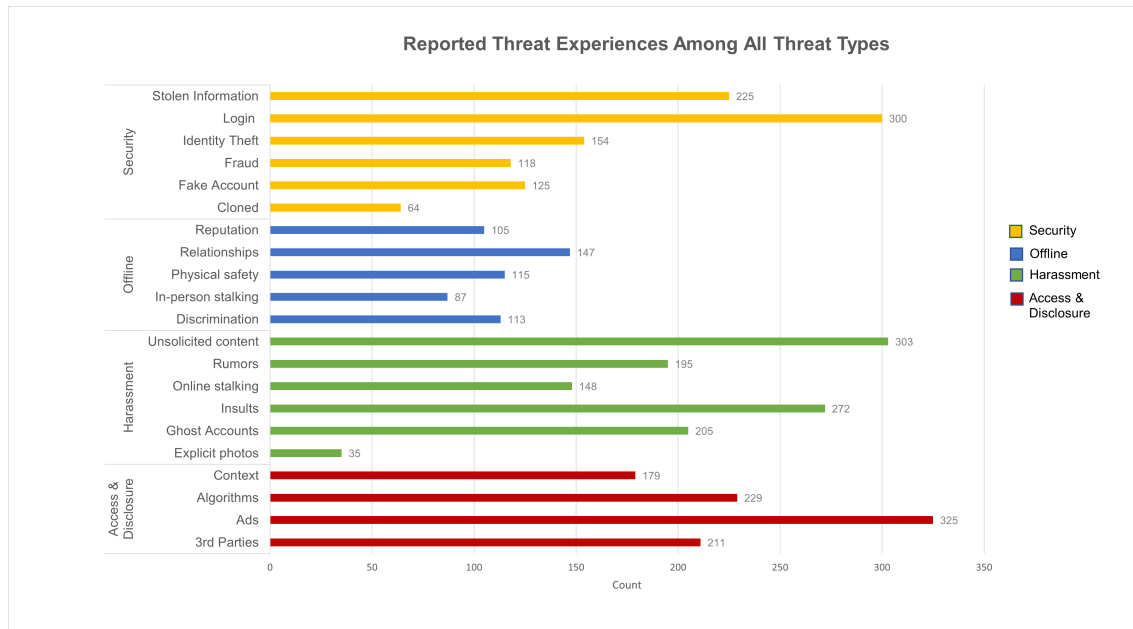


Figure 3.2: Reported prior victimization counts across all participants (N=551) and all observed threat categories.

### 3.4.2 RQ2: Which threats are perceived to be the most concerning?

We operationalize concern by examining responses related to how participants' conceptualize threats. Prior work has argued that understanding which types of experiences are perceived to be most threatening could assist in the prioritization of resource deployment for the development of

Category	Threat	Antigua and Barbuda	Barbados	Dominica	Grenada	Jamaica	Saint Kitts & Nevis	Saint Lucia	Saint Vincent	Trinidad & Tobago
Security	Identity Theft	0.63	-0.90	0.64	0.62	0.53	-0.58	-0.63	-0.53	1.01
Security	Fraud	0.80	-0.26	0.79	-1.25	0.14	0.02	-0.41	0.91	-1.51
Security	Login	-0.36	1.00	-0.62	-0.02	0.46	-0.29	0.19	-1.54	0.09
Security	Fake Account	1.69	1.68	-0.47	-1.47	-1.12	-1.06	1.72	-0.22	-0.73
Security	Stolen Information	-0.18	0.12	-0.93	0.42	-0.31	0.09	0.42	0.66	0.16
Security	Cloned	0.89	-1.62	2.26	-0.48	0.97	-1.46	-0.81	1.42	-0.09
Access	3rd Parties	-0.37	0.46	-1.23	1.08	0.52	0.06	0.30	-1.37	-0.21
Access	Ads	-0.45	0.51	0.04	-0.16	0.55	-0.59	-0.41	-0.86	1.06
Access	Algorithms	-1.08	0.96	-0.74	0.75	0.09	-0.35	-0.08	-0.61	1.31
Access	Context	0.11	1.24	-1.60	-1.00	-0.06	0.80	-0.40	-0.65	-0.43
Harassment	Rumors	0.27	-0.45	-0.53	-0.07	-0.92	1.91	-0.25	-0.06	-0.32
Harassment	Explicit photos	1.45	-2.15	1.63	-1.73	1.28	-0.77	-2.69	0.97	2.53
Harassment	Insults	-0.53	0.65	0.56	-0.06	-0.65	0.83	0.29	-0.89	-0.80
Harassment	Ghost Accounts	-0.56	0.15	1.20	-0.12	0.36	-0.54	0.74	-1.06	-0.47
Harassment	Unsolicited content	-1.60	0.00	-0.20	-0.82	2.13	-0.42	-0.14	-1.20	-0.13
Harassment	Online stalking	-0.18	0.17	-0.45	1.76	0.35	-1.00	1.14	0.34	-1.44
Offline	Discrimination	-1.46	0.89	0.84	1.03	0.96	-1.14	-1.52	0.13	0.41
Offline	Reputation	1.96	-1.59	-0.52	0.59	-2.04	1.89	0.16	0.95	0.49
Offline	Relationships	0.68	-0.20	-1.25	-0.65	-1.48	1.65	1.49	1.10	-1.58
Offline	Physical safety	-1.30	0.30	0.00	1.99	-1.08	1.09	-0.26	0.57	-0.17
Offline	In-person stalking	-0.40	-0.97	0.57	-0.40	-0.69	-0.16	1.14	1.93	0.82

Figure 3.3: Regional victimization trends. Numbers shown represent the standardized residuals. Color gradient corresponds to the magnitude of the discrepancy (Red is smaller than expected; Green is larger than expected)

protective mechanisms [86], and to better understand nuances around how protective strategies should be deployed. Thus, to assess concern, we consider patterns related to perceptions of how severe a threat is and the extent to which participants perceived themselves to be vulnerable to those threats.

Across threat categories, there were similar levels of agreement regarding which types of threats were perceived to be most severe (see Figure 3.4 and Figure 3.5 for a breakdown across different threats).

It can be referred from data in Figure 3.5 that, compared to the severity levels displayed in Figure 3.4, participants felt they were less susceptible to risks even if they considered them to be severe. This was evident for online-offline threats where participants felt it was more unlikely that they would have those experiences. In contrast, threats that impact the access and disclosure of private information were most prevalent, considered highly severe, and on average users felt most vulnerable to these threats. Among all threat types, one noteworthy outlier was participants' perceptions of their vulnerability to having their personal explicit content shared without their consent. Participants claimed to be much less vulnerable to this potential threat than to all other threats, with less than 20% of participants feeling at least somewhat likely to experience this. In essence, having explicit photos leaked is considered a very serious threat across the region. Although

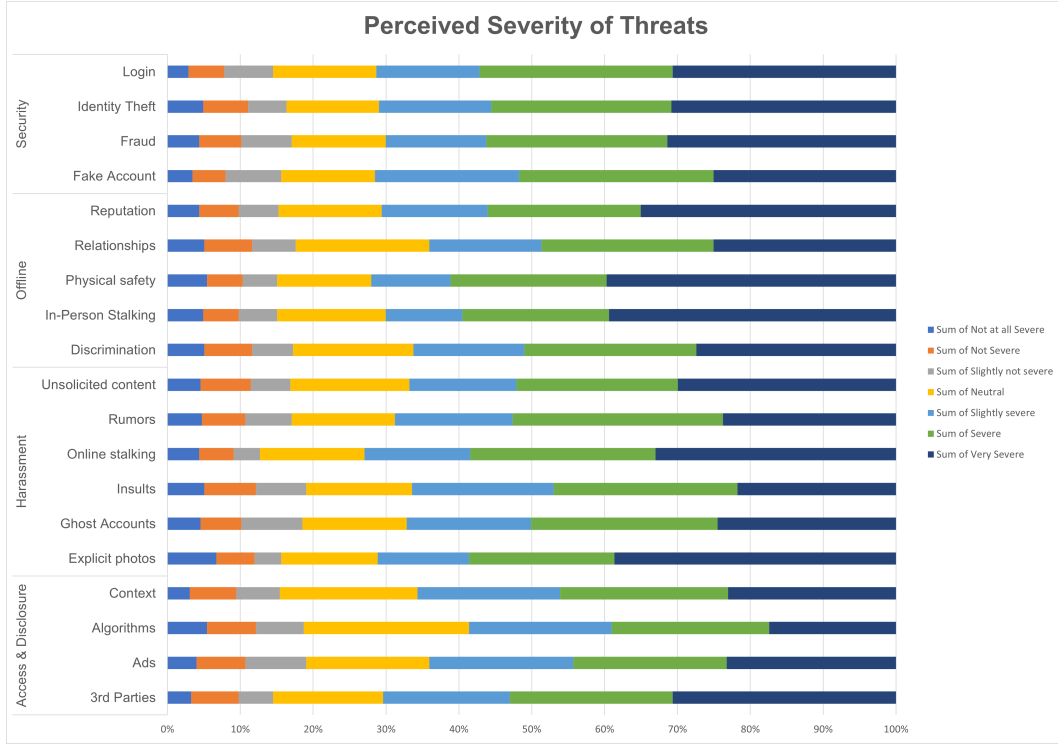


Figure 3.4: Sample-wide comparison of the perceived severity of threats across all threat categories

it is a major threat, most participants were not convinced they were likely to have that experience.

### 3.4.3 RQ3: What role does users' threat and coping appraisals play in their intention to adopt protective behaviors?

We apply structural equation modeling (SEM) to test the relationships between the PMT components, as hypothesized by theory, in four SEM models based on each type of threat—threats to digital security, threats related to access and disclosure, threats that spill over from online into offline contexts, and harassment-related threats. SEM combines confirmatory factor analysis and path analysis to test hypothesized causal relationships between latent constructs [156]. For each factor, we use multi-item measurement scales to control for measurement error [85].

To validate the robustness and validity of our measurement scales, Confirmatory Factor Analysis (CFA) was employed. Items with low loadings were removed from subsequent analyses (see the greyed-out items in Tables 1-4 in the Appendix).

Discriminant validity was assessed by comparing the average variance extracted (AVE) of

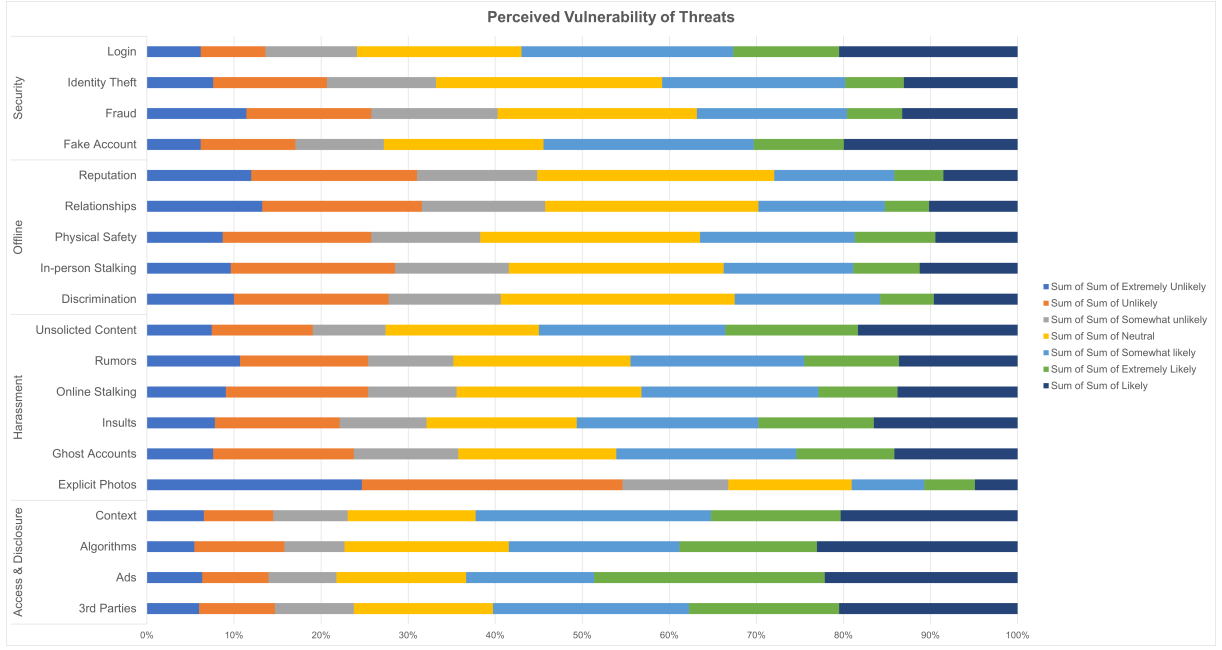


Figure 3.5: Sample-wide comparison of the perceived vulnerability of threats across all threat categories

each factor against its correlation with other factors. We found that *self efficacy* had a very high correlation with *response efficacy* in all sub-models. As such, *self efficacy* was removed from the analysis. Consequently, we do not describe results pertaining to this factor. The remaining factors exhibited a high reliability and convergent validity: Cronbach's  $\alpha$  values were excellent<sup>2</sup>, ranging between .81 and .96 while all AVE values exceeded 0.50.

We subsequently subjected the 6 factors and selected exogenous variables to Structural Equation Modeling (SEM). For the country-level analysis, we conducted omnibus tests to eliminate the possibility of family-wise errors and conducted a power analysis which confirmed that the sample sizes per country were sufficient to reveal large effects. The corresponding structural models<sup>3</sup> with the evaluation results are presented in Figures 3.6-3.9. The model fit indices for all four models indicate good to excellent fit<sup>4</sup>.

- Threats related to online-to-offline contexts: excellent fit:  $\chi^2(315) = 608.795$ ,  $p < .01$ ; RMSEA

<sup>2</sup>For alpha,  $\hat{\gamma}$  .70 is acceptable,  $\hat{\gamma}$  .80 is good,  $\hat{\gamma}$  .90 is excellent.

<sup>3</sup>Significance levels in the models are indicated as: \*\*\* $p < .001$ , \*\* $p < 0.1$ , \* $p < 0.05$ .  $R^2$  is the proportion of variance explained by the model. Numbers on the arrows represent the  $\beta$  coefficients (and the standard error) of the effect

<sup>4</sup>A model should not have a non-significant  $\chi^2$ , but this statistic is regarded as too sensitive [21]. Hu and Bentler [80] propose cutoff values for other fit indices to be: CFI > .96, TLI > .95, and RMSEA < .05, with the upper bound of its 90% CI below 0.10.

= 0.042, 90% CI: [0.037, 0.047], CFI = 0.986, TLI 0.985.

- Harassment-related threats: excellent fit:  $\chi^2(781) = 1197.256$ ,  $p < .01$ ; RMSEA = 0.032, 90% CI: [0.028, 0.035], CFI = 0.986, TLI 0.991.
- Threats to digital security: excellent fit:  $\chi^2(527) = 649.005$ ,  $p < .01$ ; RMSEA = 0.024, 90% CI: [0.019, 0.029], CFI = 0.993, TLI 0.995.
- Threats to the access and disclosure of personal information: excellent fit:  $\chi^2(517) = 814.834$ ,  $p < .01$ ; RMSEA = 0.033, 90% CI: [0.028, 0.037], CFI = 0.998, TLI 0.999.

Results pertinent to the proposed hypotheses are depicted in Table 3.3. For clarity, we report significant direct effects from left to right and endogenous variable are not depicted.

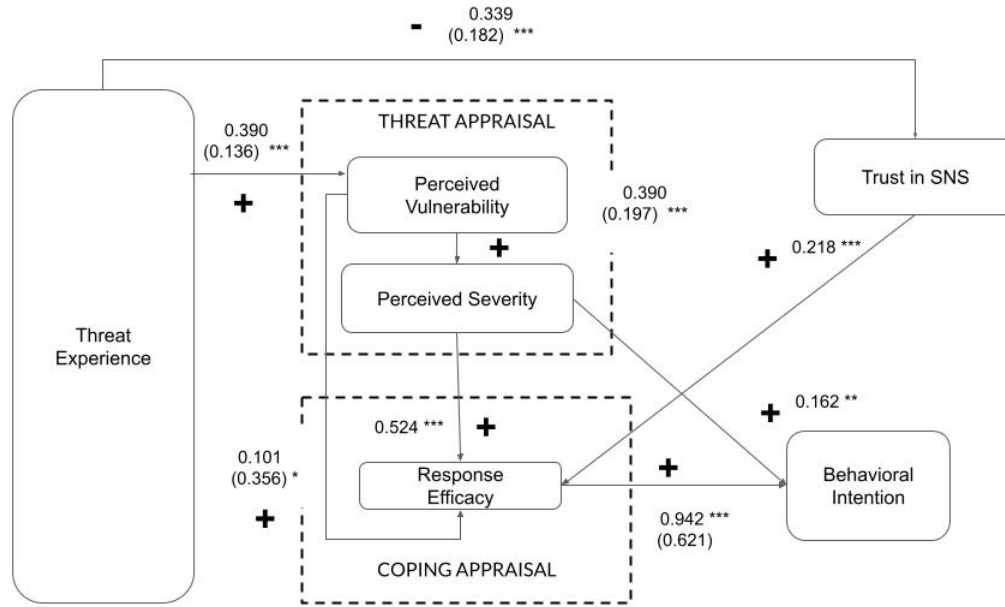


Figure 3.6: The figure above displays the SEM models for threats related to digital security

**Hypothesis 1** postulated that prior victimization would affect participants' perceived vulnerability to threats. Indeed, across all models, threat experience (i.e., prior victimization) significantly increased perceived vulnerability to threats related to: harassment ( $\beta = 0.571$ ,  $p \leq 0.001$ ), digital security ( $\beta = 0.390$ ,  $p \leq 0.001$ ), access & disclosure ( $\beta = 0.672$ ,  $p \leq 0.001$ ), and online-to-offline contexts ( $\beta = 0.583$ ,  $p \leq 0.001$ ). Therefore, H1 is supported.

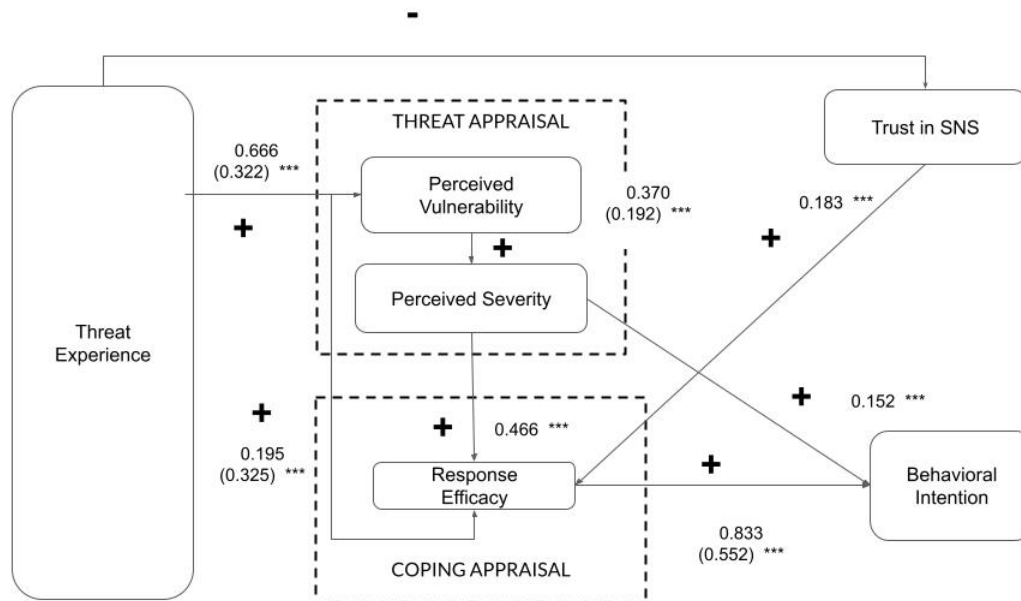


Figure 3.7: The figure above displays the SEM models for threats related to Access and Disclosure

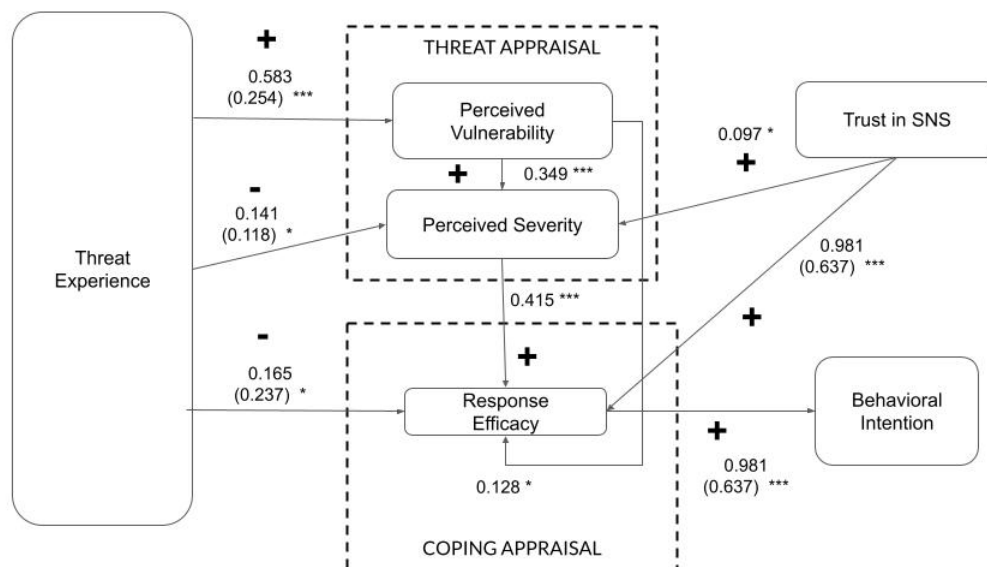


Figure 3.8: The figure above displays the SEM models for threats related to online-offline contexts

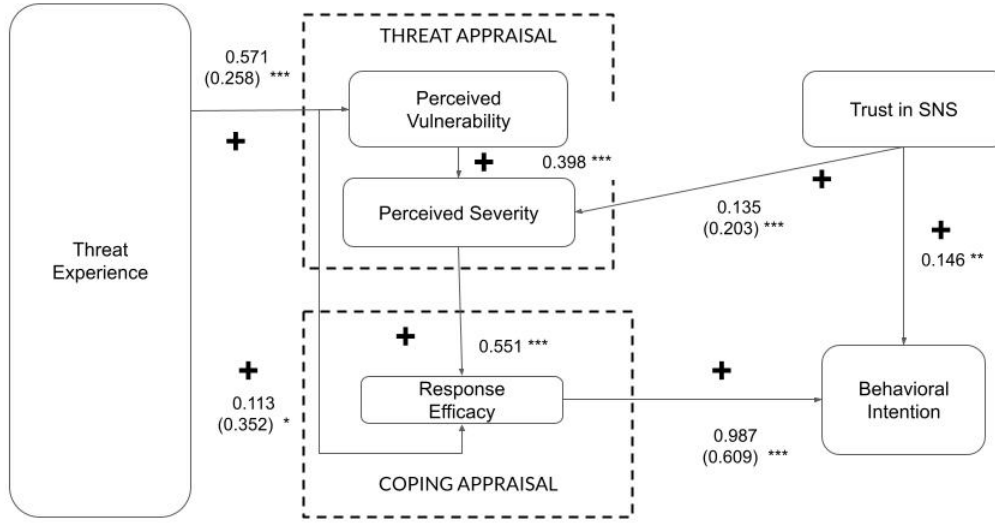


Figure 3.9: The figure above displays the SEM models for Harassment-related threats

Similarly, **Hypothesis 2** postulated that prior victimization would affect participants' perceptions of threat severity. Threat experience did not have a significant effect across all threats, except for online-to-offline threats (see Figure 3.8). In that context, there was a significant negative effect of prior threat experience on perceived severity ( $\beta = -0.141$   $p < 0.05$ ). Thus, this hypothesis is only supported in the model for online-to-offline threats.

That said, we also found a consistent significant positive relationship between perceived vulnerability and perceived severity—participants who considered themselves more vulnerable to a certain threat also considered the threat to be more severe. Consequently, while we only find a significant direct relationship between threat experience and perceived severity in the online-to-offline threat context, our models consistently show an indirect effect of threat experience on perceived severity, mediated by perceived vulnerability (i.e., participants with prior threat experience considered themselves to be more vulnerable to those threats, and subsequently perceived these threats to be more severe).

We also note a key difference in perceived threat severity across countries. Figures 3.13, 3.10, 3.11, 3.12 provide an overview of the differences in perceived severity by country. Notably, participants from St. Lucia reported higher levels of perceived severity across all types of threats.

Comparatively, St. Lucian participants perceive threats related to digital security ( $\beta = 0.617$ ,  $p < 0.01$ ) and access and disclosure of personal information ( $\beta = 0.482$ ,  $p < 0.05$ ) at a significantly higher level of severity.

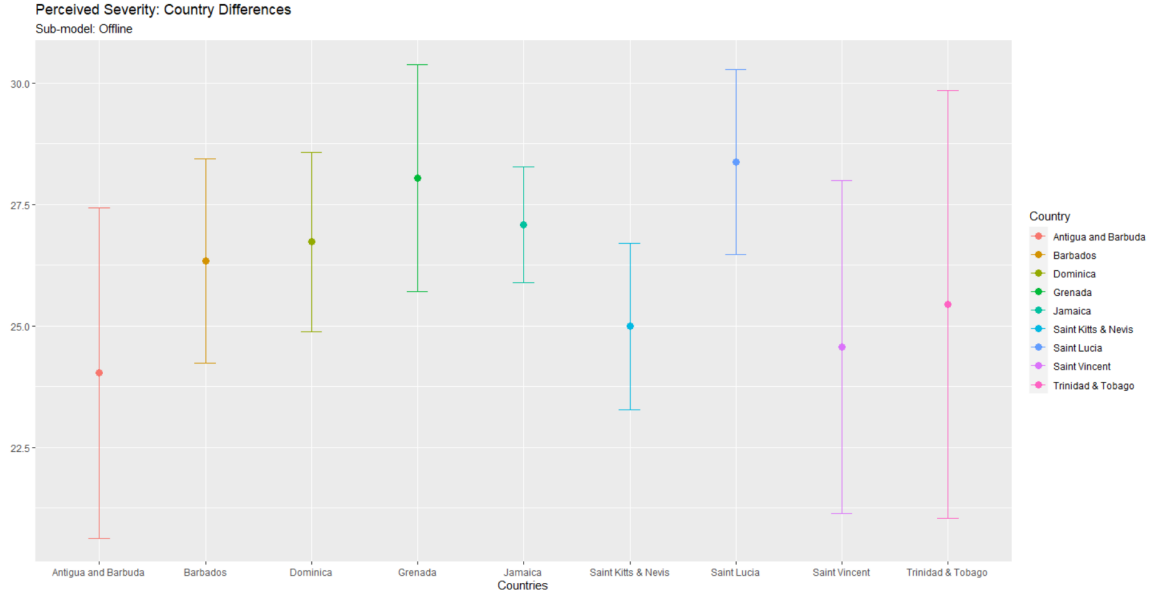


Figure 3.10: Marginal effects of perceived severity for online-to-offline threats

**Hypothesis 3** postulated that prior victimization would affect participants' perceptions of trust in social media platforms. Participants who had a higher level exposure to threats had more negative attitudes regarding the trustworthiness of social media platforms. Trust in social media platforms significantly decreased as participants had experiences with threats related to harassment ( $\beta = 0.571$ ,  $p < 0.001$ ), digital security ( $\beta = 0.390$ ,  $p < 0.001$ ), access & disclosure ( $\beta = 0.672$ ,  $p < 0.001$ ), and online-to-offline contexts ( $\beta = 0.583$ ,  $p < 0.001$ ). This provides supporting evidence for H3 in all models.

**Hypothesis 4** postulated that prior victimization would affect participants' coping appraisal. This effect was only significant for online-to-offline threats (see Figure 3.8), where prior victimization negatively impacted the extent to which participants felt safety tools would help them to remain safe ( $\beta = -0.165$ ,  $p < 0.05$ ).

That said, we also found consistent significant positive relationships between perceived vulnerability / severity and response efficacy—participants who considered themselves more vulnerable to certain threat and who considered these threats to be more severe also felt that safety tools



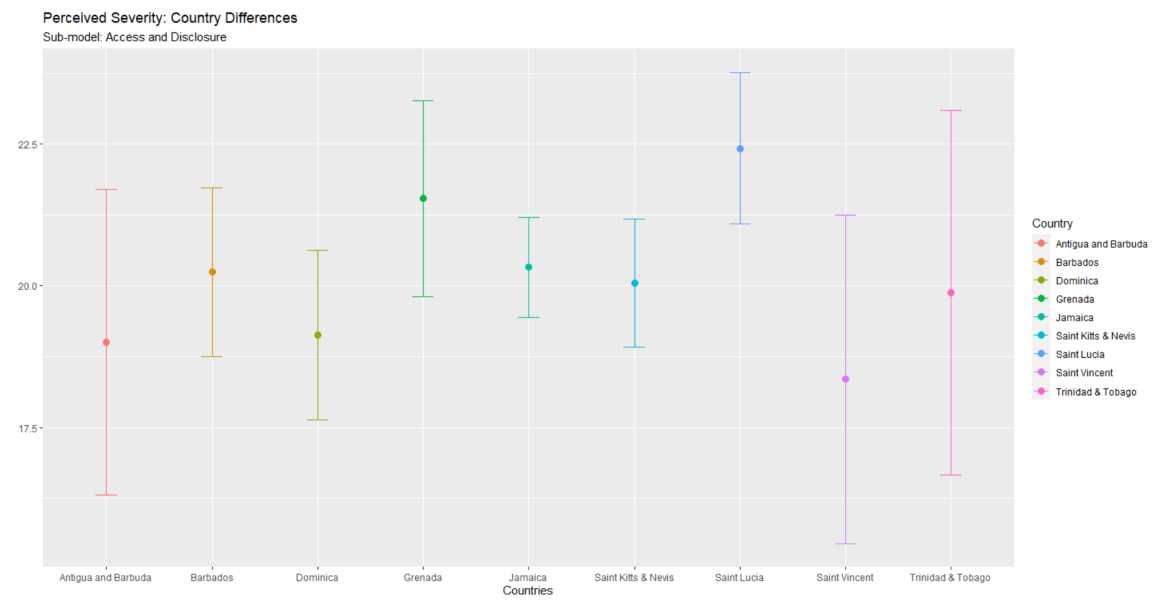


Figure 3.11: Marginal effects of perceived severity for threats related to Access and Disclosure

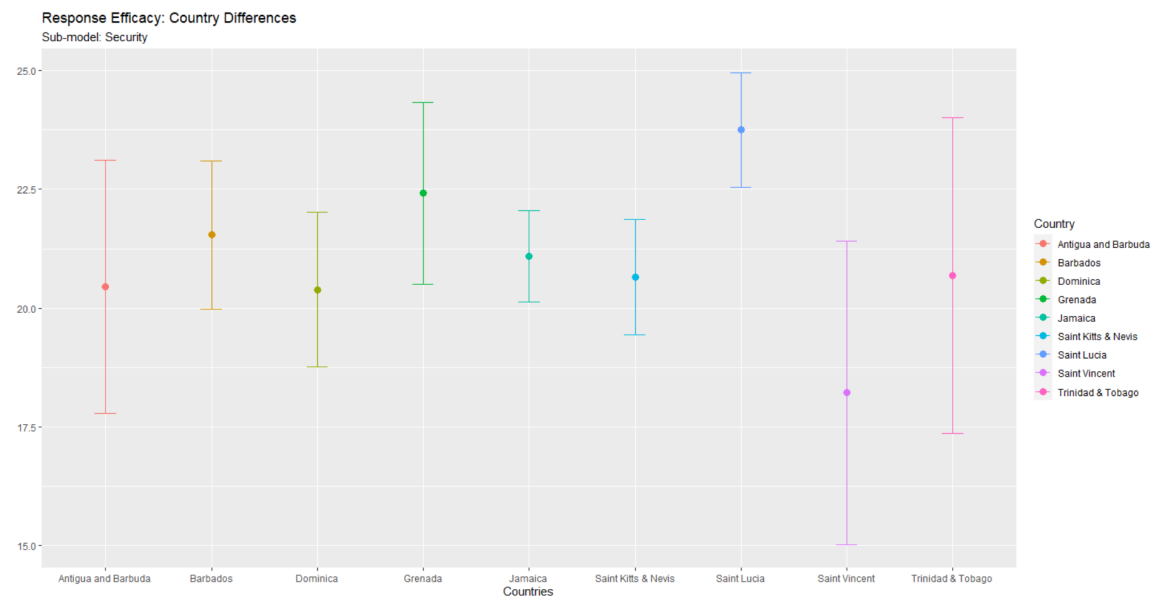


Figure 3.12: Marginal effects of perceived severity for security threats

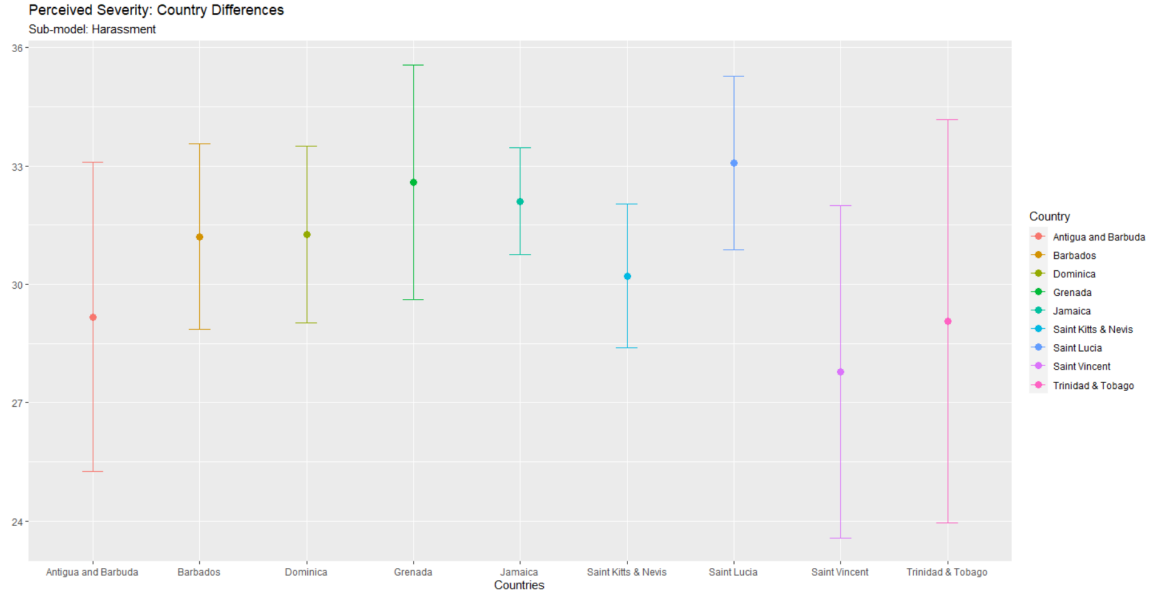


Figure 3.13: Marginal effects of perceived severity for harassment-related threats

would help them remain safe. These effects can be explained if one considers that people who feel vulnerable towards severe threats are likely to expend more effort familiarizing themselves with potential protective behaviors. This familiarity could then increase their confidence in responding to the threat (cf. [22]). Consequently, while we only find a significant direct relationship between threat experience and response efficacy in the online-to-offline threat context, our models consistently show an indirect effect of threat experience on response efficacy, mediated by perceived vulnerability and perceived severity.

**Hypotheses 5, 6 and 7** postulated that resp. users' coping appraisal, perceived severity and perceived vulnerability influenced their behavioral intention to implement protective behaviors. Among these, only the relationship between response efficacy and behavioral intention was consistently found to be significant, supporting H5. Participants who felt that safety tools would help them to remain safe had a higher intention to adopt behaviors to prevent threats related to harassment ( $\beta = 0.987$ ,  $p < 0.001$ ), digital security ( $\beta = 0.943$ ,  $p < 0.001$ ), access & disclosure ( $\beta = 0.833$ ,  $p < 0.001$ ), and online-to-offline contexts ( $\beta = 0.981$ ,  $p < 0.001$ ).

Participants' perceptions of vulnerability were associated with their behavioral intention to implement protective behaviors as well, but this effect was not consistent across the four threat categories. Participants who perceived higher levels of severity were more likely to adopt protective

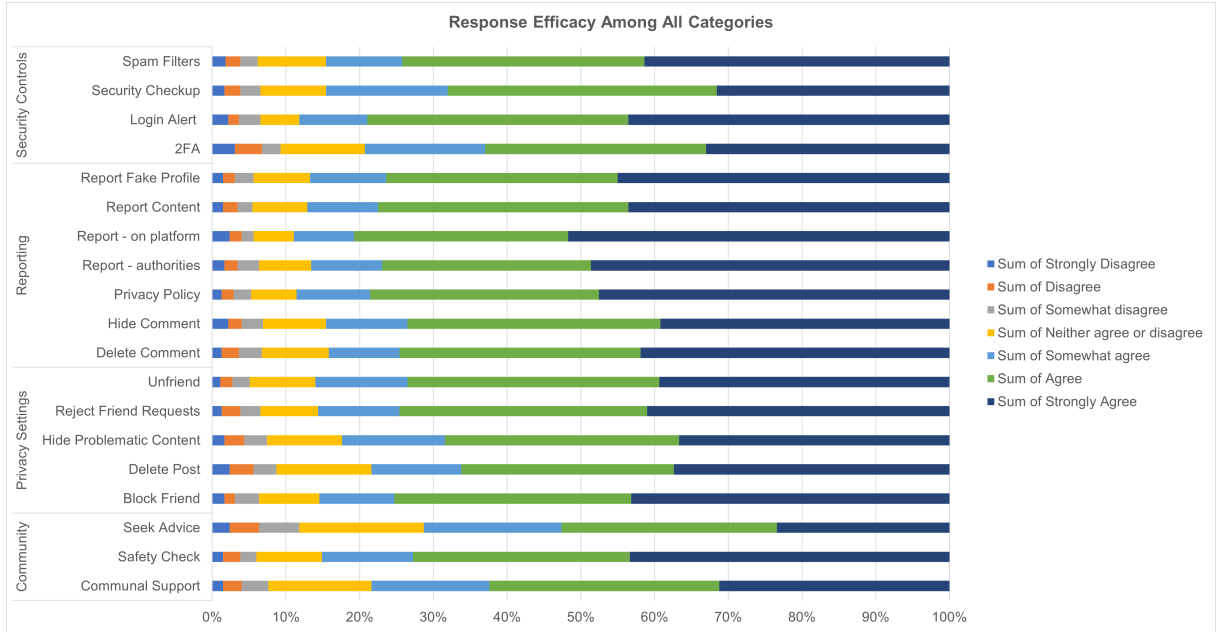


Figure 3.14: Sample-wide comparison of the response efficacy across all protective behavior categories

behaviors for threats related to their digital security ( $\beta = 0.162, p \leq 0.01$ ) and access and disclosure of their personal data ( $\beta = 0.152, p \leq 0.01$ ). Thus, H6 is only supported for these two types of threat. In contrast, we found no significant associations between perceived vulnerability and behavioral intention. As such, H7 is not supported.

We note, though, that due to the effect of response efficacy on behavioral intention and the effects of threat appraisal on response efficacy, perceived vulnerability and severity do have an *indirect* effect on behavioral intention, mediated by response efficacy. In other words, a high threat appraisal likely caused users to increase their response efficacy (e.g., by familiarizing themselves with potential response strategies), which in turn increased their intention to implement protective behaviors.

Lastly, we conducted tests to investigate associations between participants' trust in social media platforms and their intention to adopt protective behaviors. Our results reveal that trust only played a significant role in the adoption of protective behavior for harassment-related threats ( $\beta = 0.146, p \leq 0.01$ ). We also note that trust in social media platforms increased users' response efficacy in all models except the model for harassment-related threats. Arguably, trustworthy social networks can help users mitigate threats, except for harassment-related threats. Due to the serious nature of such threats, it might be worthwhile for social media platforms to consider ways to help

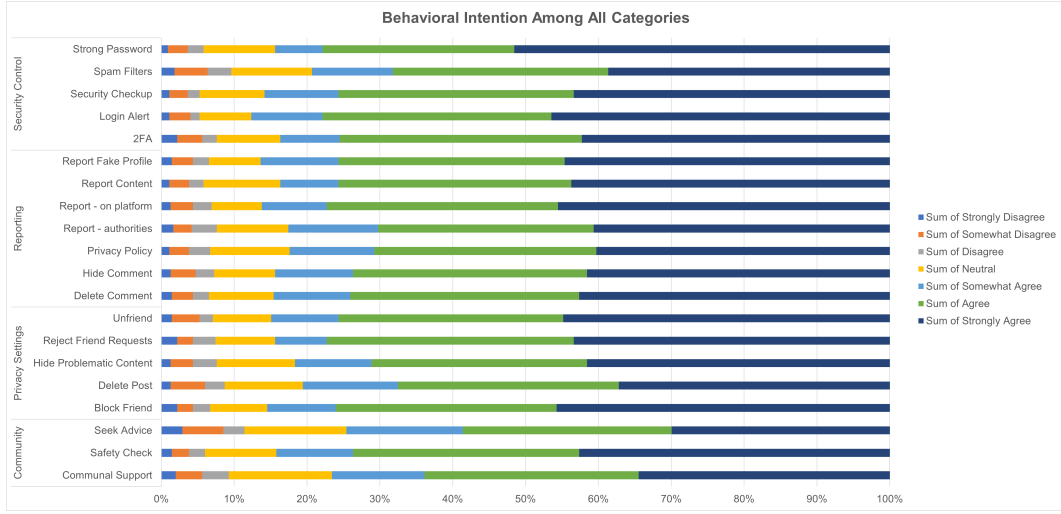


Figure 3.15: Sample-wide comparison of behavioral intention across all protective behavior categories

users increase their response efficacy against them.

We summarize the findings in tables 3.3 and 3.4.

## 3.5 Discussion

In this section, we discuss emerging insights and theoretical implications based on our data about the perceptions of threats and their impact on protective behaviors. We develop our understanding through interviews with local experts from a diversity of backgrounds, who helped us contextualize the results and ensured that the implications are reflective of the needs of people in the region. We discuss practical design and policy implications of our study. We close by presenting our research limitations and directions for future work.

### 3.5.1 Theoretical and Practical Considerations

Overall, our findings illustrate that there are significant variations across different countries in how people evaluate threats, and our model shows that these differences subsequently influence their safety intentions. This provides supporting evidence for the perspective that challenges the one-size-fits-all approach to safety mitigation currently employed by social media platforms.

Despite these variations, there are similarities in perceptions in the underlying factors that influence social media users' intention to protect themselves against threats to their safety. First,

Hypothesis	Description	Access	Security	Harassment	Offline
H1	Threat experience will have an effect on perceived vulnerability	Supported	Supported	Supported	Supported
H2	Threat experience will have an effect on perceived severity	Partially Supported	Partially Supported	Partially Supported	Supported
H3	Threat experience will have an effect on perceptions of trust in social media platforms	Supported	Supported	Not Supported	Not Supported
H4	Threat experience will have an effect on users' coping appraisal	Not Supported	Not Supported	Not Supported	Supported*
H5	Users' coping appraisal is positively associated with behavioral intention	Supported*	Supported*	Supported*	Supported*
H6	Perceived severity is positively associated with behavioral intention	Supported	Supported	Partially Supported	Partially Supported
H7	Perceived vulnerability is positively associated with behavioral intention	Not Supported	Not Supported	Not Supported	Not Supported
H8	Trust in SNS is positively associated with behavioral intention	Partially Supported	Partially Supported	Supported	Partially Supported

Table 3.3: The table above describes the summary of findings related to the hypotheses testing. Items denoted by (\*) signify hypotheses where coping appraisal, which comprises of self efficacy and response efficacy, is observed but self efficacy was dropped and the results reflected represent response efficacy only.

the results point to safety being a pervasive challenge across the region: across all countries, an overwhelming number of persons reported having encountered a threat to their safety at least once. Despite this, prior victimization increased users' motivation to protect themselves going forward. Aligning with Protection Motivation Theory [122], perceived vulnerability towards threats, perceived severity of threats, and perceived response efficacy in protecting against threats significantly contributed to users' intention to engage with protective behaviors. Unlike the original PMT model, we do not find a direct relationship between threat appraisal components and behavioral intention. Rather, we find that perceptions of severity and vulnerability influence safety intentions only via people's response efficacy.

Taken together, people who had previously faced threats perceived higher vulnerability, higher vulnerability resulted in higher perceived severity, which in turn increased users' response efficacy, which in turn increase their intentions to engage with protective behaviors. Researchers have argued that risk exposure builds resilience and aids in risk mitigation [196]. Conversely, though, this means that persons with lower levels or no experiences with threats, such as younger audiences or social media users with fewer technology skills, may initially refrain from protecting themselves—

Research Questions	Results	Implications
RQ1: Which types of threats are prevalent?	The top three threats participants experienced were related to unwanted ads, unsolicited content, and stolen login credentials. Participants from St. Vincent had the highest average incident rate across all threats.	For platform designers, creating easily accessible and actionable control options could assist in mitigating unwanted interactions. More visibility of security practices could assist in reduces incidents of stolen login credentials.
RQ2: Which threats are perceived to be the most concerning?	Threats to the access, collection, and disclosure of personal information were most concerning. This threat category was most prevalent, people felt they were severe, and they felt most vulnerable to these threats. On a country-level perspective, participants from St. Lucia were most concerned about experiencing threats overall.	High incident rates coupled with high rates of perceived vulnerability may indicate either a need for better awareness of existing tools or a need for tailored tools for more protection.
RQ3: What role does users' threat and coping appraisals play in their intention to adopt protective behaviors?	For harassment and online-offline threats, people are willing to adopt protective behaviors depending on how well they think protective measures actually work regardless of the severity. In comparison, the severity of the threat plays a direct role in using protective measures for threats related to security, and access and disclosure.	There might be gaps in the effectiveness of protective measures for harassment and online-offline threats. Users would experience more severe threats and only be motivated to use measures based on how well they think the platform would help.

Table 3.4: In the table above, we summarize the results in relation to our research question as well as offering an overview of the respective implications.

until they are victimized. It could be distressing for victims who may encounter risks for the first time and who may not be completely aware of what to do.

This could be exacerbated for younger social media users who may ask their parents for support, but their parents do not understand the threat itself or be unfamiliar with available options for redress. For example, discussions with experts revealed that there have been multiple instances of severe consequences for high school students in Trinidad, where the creation of malicious explicit deepfakes have been rampant lately. The expert explained

*“because of a lack of knowledge in terms of what technology brings to the table and the kind of things that could happen it was difficult for him [the parent of the victim] initially to accept that this [the deepfake] wasn’t really happening. It was only when the daughter attempted suicide, that the family decided to seek help”* — E6, Director of a non-profit organization, Trinidad and Tobago.

Adopting an approach that encourages resilience through risk exposure would be impractical:

The consequences of exposure to high-level risks are severe and when that severe risk is coupled with the continual evolution of threats, it raises questions about the long-term effectiveness of such a reactive safety mitigation strategy. Therefore, despite safety intentions being increased by prior victimization, exposure should not be central in mitigation approaches, as the consequences of negative experiences could be irreparable.

Generally, there was agreement regarding the severity of harms. Regionally, threat appraisal was high: Caribbean people felt that threats in all categories were severe and that it was not unlikely for them to personally encounter such threats. Notably, among all threats, the highest reported risk was being sent unsolicited content. However, most persons thought there was a very low likelihood that they would ever experience their own explicit photos being shared without their consent. This is of particular interest since there have been multiple media reports across the region of women and girls being exploited and harassed by men who unbeknownst to them shared their explicit photos [11, 67, 132]. Upon further investigation of these media reports, we note that the majority of the perpetrators were persons with whom the victims had close ties (e.g. domestic partners or friends). Therefore, a possible explanation for the discrepancy between threat exposure and vulnerability might be that persons initially do not expect close ties to violate boundaries regarding content they feel protected by co-ownership. This is consistent with Petroni’s Communication Privacy Management theory (CPM) which explains that people have heavily guarded boundaries for private content and thus anyone who has access to that information should treat the content in the same regard [141]. Relationships change, though, and the potential adversarial nature of a break-up can threaten to disrupt these heavily guarded boundaries. To mitigate harms in such situations, designers should consider intuitive and fail-safe means to revoke co-ownership of intimate content between (ex-)partners.

One of the contributions of this work is the inclusion of threats with offline consequences that occur as a result of online interactions. Close knit societies like the Caribbean are more integrated, and thus the perceptions of severity for such offline threats may differ from typical WEIRD societies. Previous studies have illustrated that cultural norms serve as a significant predictor of online disclosure [110, 103]. As such, social media users from individualist cultures may have safety concerns centered around how the consequences of risk exposure will affect them personally, while users from collectivist cultures like the Caribbean may be more concerned about the collective consequences of their risk exposure for their strong ties (e.g. friends and family) [181]. As representative

proponents of this view, our experts described:

*"it is not easy to recover here. Let's say you were living in New York. How many people actually know you there? Here, if your character is assassinated online, even if it true or not, that is ingrained in the minds of everyone. Then you have to consider how this will affect those around you. How that will affect your options for jobs and options for your family members."* - E1, Youth Ambassador, St. Kitts-Nevis.

Therefore, we encourage further research to explore threats that spillover into the physical world and other diversely perceived and complex harms.

Furthermore, our findings highlight that perceptions related to the efficacy of safety tools are central to users' intention to engage in protective behaviors irrespective of the type of harm. This would be critical for stakeholders to consider when designing options for redress: If people are expected to adopt mitigation methods, there should be enough transparency about the effectiveness of the available tools to inform their safety decision-making process.

### 3.5.2 Design and Policy Implications

The results in this paper provide numerous opportunities to build upon and deepen the current body of knowledge surrounding online safety for the HCI community and beyond. First, the design of many of the safety mechanisms offered to social media users focuses on *equality*: All platform users are afforded the same resources and opportunities for risk mitigation. While this is an admirable endeavor, it fails to acknowledge that giving the same resources does not lead to the same outcomes for those who may be disproportionately disenfranchised by for imbalances, inequalities, and injustices. To illustrate, we observed that Caribbean people were just as motivated to engage with reporting tools on the platform as they were with offline reporting options (e.g. building a legal case) even though a considerable number of countries in the region do not have substantive laws for redress in case of online harms [135]. In light of this, we encourage platform owners to adopt an approach grounded in *equity* rather than equality, which would uncover the appropriate resources needed to elevate the positions of disenfranchised users, so as to achieve fair outcomes for all users. For developers and designers, this would require going a step beyond the "one-size-fits-all" approach to online safety and ensuring that the resources are accessible and effective. For example, this could involve lobbying for the establishment of local online safety laws, so that the platform's reporting



tools can indeed be used to seek legal redress. This aligns with recent work that has advocated for platforms to integrate a tailored "constitutional layer" that is responsive to local context [28]. Thereby, future AI-enabled tools could assist victims in retrieving potential supporting evidence from their devices (such as call logs, messages, summary reports of interactions) to assist in making reports or preparing for a legal case. This option would be helpful for regions with similar pain-points as the Caribbean where there might not be widespread access to information about the procedures of justice. Outside of the region, the concept of equitable design in privacy and safety could be applied to marginalized groups in Western countries to assist with offering additional support or proving easier access to tools that would help them achieve fair outcomes.

In a similar vein, our findings and input from local experts raise concerns about a reliance on reactive justice. Across the Caribbean, there are threats that impede people's ability to safely use the internet while many are concerned about the impacts of post-digital threats lingering from their online interactions. On a platform-level, tools are tailored for retributive justice while more culturally-appropriate options such as mediation are not implemented. Many justice-oriented techniques rely on exposure to harms (e.g. problematic online content), since the success of these approaches depend on users reporting the harms (e.g. flagging the content). Instead, we support a new direction of alternative approaches to justice that depart from solely punitive techniques (e.g. banning users). Along these lines, Schoenebeck and Blackwell argue that social media governance has revolved around Western models of criminal justice, which is centered on compliance with formal rules versus the accountability for and repair of specific harms [157]. The results from our study suggest that Caribbean internet users are experiencing threats that trample their basic human rights to preserve their privacy and safety as individuals. Thus, heavily utilizing reactive models comes at the cost of overburdening millions while malicious actors prevail. Regionally, collective efforts to implement and deploy proactive technological tools might prove to be financially straining and logistically draining since many countries have varying priorities for their limited resources. To combat this, we suggest a combined effort

Lastly, our analysis revealed that individuals who are geographically co-located may still display distinctive views, which undoubtedly has implications for regional legislation. The results point to countries that might need to devote additional resources to education to encourage the adoption of protective measures or education campaigns to ensure people are aware of the rights to safety online. For example, CARICOM (an intergovernmental organisation of 15 member states

throughout the Caribbean) has recently launched an initiative aimed at offering legislative protection for Internet users. Our data offers insight into the types of threats that are most prevalent, those that are perceived as most severe, and the types of strategies people throughout the region are willing to employ. Thus, the insights could help to inform policy, design, and the development of safety-related mechanisms. That said, our results also demonstrate some substantial differences within the region, suggesting that a supranational legislative approach must have ample opportunity for local nuances and adjustments.

### 3.6 Chapter Conclusion

This study offered empirical evidence about non-Western social media users' motivations for adopting behaviors that protect them against pervasive threats to their privacy, security, and personal well-being. While research to date on specific types of harms have been siloed, we offer a holistic view on how people in a non-Western context perceive and evaluate online threats. Moreover, by conceptually defining protective behaviors based on the threats that they address, we were able to build knowledge on how the perceptions of threats influence the adoption of online safety mechanisms. The study uncovered nuanced differences among threats related to harassment, digital security, access and disclosure, and online-to-offline threats—as well as between different countries in the Caribbean. These findings offer several contributions:

- We build on existing Human Computer Interaction (HCI) theory by presenting a conceptual model for engaging HCI researchers, designers, and policy advocates in online safety research.
- To the best of our knowledge, this study is the first to conduct a regional survey on online safety within the Caribbean, which contributes to the limited body of existing HCI research on this population and towards knowledge on the prevalence of threats region-wide.

Our findings provide a new understanding of users' mental models, behaviors, and attitudes with respect to online safety. However, it remains clear that there are design opportunities for inclusive and equitable safety tools. In Chapter 4, I consider the findings in this chapter and investigate routes for maintaining online safety outside of platforms.

## Chapter 4

# A Critical Reflection of Legislative Protections in the Caribbean

The findings in chapter 3 provide an important overview of the threats Caribbean citizens face and their attitudes towards technical interventions to address them. In particular, people are less likely to adopt technical interventions to address interpersonal harms such as harassment if they do not trust the platform. This then raises questions about how persons seek fair outcomes when technical solutions are not deemed trustworthy or effective. Recently, advocates have warned against “technological solutionism” as the default option for responding to harms that were facilitated by technological systems to begin with [159].

Fairness in algorithmic and data intensive systems are globally recognized as topics of critical importance and protections exist within and outside of the technological realm to uphold the principles of fairness. Regulators have recognized the importance of protections and multiple countries have enforced laws that directly address offences facilitated and amplified by technology such as revenge pornography, defamation, harassment, and cyber-stalking [140, 154, 44, 37, 91]. The entangled nature of ADDTs has wrapped technological problems with existing societal ones. Yet, to date scholars have primarily focused on US or EU regulatory perspectives [188, 131, 202, 71, 17, 95]. More recently, researchers within the Caribbean have provided a limited scope on legislative approaches to online safety in the region [14, 13, 53]. Barclay offered a critical review of the definitions and applications of the legal protections for Jamaican citizens, however, this analysis was restricted

to Jamaica [13]. Donald et al. adopt a wider scope to legal protections in the global and south which included over 20 countries across the region but the analysis does not offer readers insight into which offenses are covered [53].

The primary objective of this chapter is to critically reflect on existing orientations towards criminal justice and punitive sanctions that protect Caribbean citizens in online spaces. Considering the aversion to the adoption of technical protections uncovered in chapter 3, this work seeks to answer the following main question: *if Caribbean citizens encounter online threats, what regulatory protections are available?*

To address this question, this chapter reports findings from a mixed methodological approach: a comparative legal analysis of relevant regulatory protections and a content analysis of media reporting of online threats. We address our main question by conducting a critical review of substantive law provisions across the region to identify patterns of harmonization and areas of discrepancies. We contextualize and confirm emerging norms from the legal review through media reports across the region. The analysis showed that regulatory approaches to online safety in the Caribbean is fractured and additional procedures are needed to promote a sustainable path for achieving justice from online offenses through judicial means. Thus, the key contribution of this chapter is a landscape assessment specifically focused on English-speaking CARICOM member states (13 countries in total). We argue that this contribution could be beneficial in understanding the gaps that exist and the hurdles that prevent a more consistent and effective approach to online safety regulations and enforcement. Moreover, this work aims to be a fundamental step towards the development of actionable principles that acknowledge geopolitical, societal, and legal influences. As such, we envision this landscape assessment to serve in a practical capacity by helping to inform regulatory agencies in prioritizing alternative conceptualizations and policy options when considering new regulations. We hope that our contribution can assist in providing supportive evidence towards understanding what is currently done, what aspects might not be serving people well, and what areas could be improved.

In the following sections, we provide: a summary of the historical influences of legislative models in the region, a review of the methodology employed in the study, a discussion of the common protections across the region and a presentation of the legal protections in each country. Next, we offer insights from our review of media reports across the Caribbean focused on online threats. Lastly, in the discussion section we offer insights and recommendations for future work.

## 4.1 Legislative History

The criminal approaches to tackling cybercrime (criminal violations that occur in online spaces or facilitated by technology) vary greatly across different countries. This heterogeneity in the interpretation and application of cybercrime law poses challenges in sustainable considerations for fairness across the region. For a united approach towards cybercrime legislation, all members of Caribbean Community and Common Market (CARICOM) signed to the Harmonization of ICT Policies, Legislation, and Regulatory Procedures (HIPCAR) project and the Budapest Convention. The aim of their participation was to build common ground on the elements that should be in a united cybersecurity strategy in response to the increasing threats of cybercrime.

” A *cybersecurity* strategy can be compared to fighting burglary by installing locks on a front door and surveillance cameras, whereas a *cybercrime strategy* can be compared to ensuring that law enforcement has the capacity to catch the thieves if they commit a crime. It would be impossible for law enforcement to deal with theft if there was not adequate technical security such as strong front doors to protect valuables. However, security features alone will not prevent crime either, there’s a need to be a legal sanction and not just technical protection to act as a deterrent to commit crimes. [135]”

Both the HIPCAR and the Budapest Convention have had considerable influence on the development of cybercrime legislation in CARICOM countries. However, both models face shortcomings. Clough argues that the language used fails to address the nature of new technological threats such as identity theft, grooming, and spam [40]. Regardless, these models have had significant impact on the development of cybercrime laws in the region. There are three main bodies of legislation that specifically address cybercrimes across CARICOM member states: the Computer Misuse Act, Electronics Crimes Act, and the Cybercrimes Act. Donalds et al. draw comparisons across the models and identify that the Computer Misuse laws across the region have been influenced by Commonwealth countries such as the UK [53] while Electronic Crimes Acts displayed strong influences from the HIPCAR model, and countries by Cybercrimes Acts aligning with provisions from the Budapest Convention.

In this chapter, we produce a comparative report detailing principles underpinning cyber-crime legislation as an approach to online safety. taken to hate crime in a range of jurisdictions. The section below outlines of objective of this research.

### 4.1.1 Objectives

The study aims to understand the legal frameworks that are in place for the protection of citizens across a range of jurisdictions. The aim of this chapter is to draw a critical review of the key legislation in the Caribbean regarding the prevention and prohibition of online abuses. Specifically, we consider social norms, media discourse, and local practices to identify challenges that may affect the enforcement of the law. Thus, the objectives of this study include:

- Reviewing the primary bodies of legislation that criminalize violations that occur in online spaces or are facilitated by technology
- Identify areas in the operationalization of laws that may pose challenges
- Draw appropriate recommendations that would enhance the implementation of online safety policies across CARICOM member states

In the following section we outline the steps taken to address our central research question and research objectives.

## 4.2 Method

The Caribbean region is home to almost 30 sovereign countries with connections based on culture, language, geopolitical similarities, and history. It is common practice to combine efforts on major issues throughout the region for harmonious integration. Thus institutions like CARICOM are essential in driving collective regulatory approaches that will impact millions of people throughout its member states. Therefore, for the purposes of this study we consider English speaking CARICOM member states as our inclusion criteria.

To accomplish the objectives of this study, we collected the respective pieces of legislation in the respective countries. We consulted with two practicing barristers to ensure we were (1) provided adequate access to the needed legislation and (2) offering a representative snapshot of the legal statutes available in the respective countries. In total, 13 countries were included in the final analysis. These include The Bahamas, Belize, Jamaica, St. Kitts and Nevis, Antigua and Barbuda, Montserrat, Dominica, St. Vincent and the Grenadines, St. Lucia, Grenada, Barbados, Trinidad and Tobago, and Guyana.

After collecting all the relevant legislation, we summarized the substantive law provisions across the corpus for country to country comparison. Offenses were categorized based on the nature of the harm and the target. Three categories emerged:

- *Computer-related offenses*: these are offenses where the target is the technology (i.e. a device), access to data, or fraudulently seeking information via electronic means.
- *Content-related offenses*: these are offenses where content is manipulated to cause psychological harm or the distribution of the content is a violation of expectations of personal privacy.
- *Interactional offenses*: these are offenses where interaction facilitated through technology caused harm.

At the conclusion of the legal analysis, we collected 127 news articles across the 13 countries. The goal of this effort was to develop an understanding of how stories of online harm were reported in the media. We present descriptive insights to contextualize the results from the legal analysis.

## 4.3 Findings

In this section, we organize the results of our legal analysis by first summarizing the substantive law provisions in the region. Next, we present an overview of the status of legal protections that are enforced with the aim of highlighting the region’s current standing regarding the extent of available laws from country to country. Lastly, we present individual country profiles which offer a synopsis of each country’s regulatory approach to online safety.

### 4.3.1 Substantive Law Provisions

The list below summarizes the range of offenses that threaten online safety and are punishable by law:

1. *Illegal access*: the intention is to cover the basic offence of dangerous threats to and attacks against the security (i.e. the confidentiality, integrity and availability) of computer systems and data.

2. *Illegal interception*: the intention is to protect the right of privacy of data communication. The offence represents the same violation of the privacy of communications as traditional tapping and recording of oral telephone conversations between persons.
3. *Data interference*: the aim is to provide computer data and computer programs with protection similar to that enjoyed by corporeal objects against intentional infliction of damage. The protected legal interest here is the integrity and the proper functioning or use of stored computer data or computer programs.
4. *System interference*: the aim is to criminalize the intentional hindering of the lawful use of computer systems including telecommunications facilities by using or influencing computer data. The protected legal interest is the interest of operators and users of computer or telecommunication systems being able to have them function properly.
5. *Misuse of devices*: the aim is to criminalize the intentional commission of specific illegal acts regarding certain devices or access data to be misused for the purpose of committing the above-described offences against the confidentiality, integrity and availability of computer systems or data.
6. *Computer related forgery*: the purpose is to create an offence of the forgery of tangible documents. It aims at filling gaps in criminal law related to traditional forgery, which requires visual readability of statements, or declarations embodied in a document and which does not apply to electronically stored data. Computer-related forgery involves unauthorized creating or altering stored data so that they acquire a different evidentiary value in the course of legal transactions, which relies on the authenticity of information contained in the data, or is subject to a deception. The protected legal interest is the security and reliability of electronic data, which may have consequences for legal relations.
7. *Computer related fraud*: the intent is to criminalize any undue manipulation of data or an electronic system that is presented as genuine when it is not.
8. *Offences related to child pornography*: the goal is to strengthen protective measures for children, including their protection against sexual exploitation through any electronic system.
9. *Offences related to the non-consensual distribution of explicit imagery*: the intention is to



make it punishable to produce or distribute explicit sexual imagery or imagery of the private area of a person.

10. *Electronic terrorism:* the goal is to criminalize the premeditated attacks where the target is an electronic system or data which results in physical injury, death, or financial harm.
11. *Electronic Defamation:* the aim is to make it punishable by law to use electronic systems to defame or tarnish the character of an individual.
12. *Cyber stalking:* the goal is to criminalize the use of any electronic system to intimidate, coerce, or annoy an individual.
13. *Spam:* the objective is to offer protections against persons who use electronic systems to transmit unwanted electronic messages.
14. *Identify-related offenses:* the purpose is to provide protection against persons who utilize electronic systems to transfer or obtain the identity of another. This is also inclusive of ones personal signature or password.
15. *Child luring:* the aim is to penalize persons who use electronic systems to engage in communications with a child that takes a sexual nature, leads to sexual activity, or used to arrange a physical meeting to abuse a child. A key distinction is that this is considered an offense whether or not a physical meeting takes place.
16. *Harassment-related offenses:* the intent is to provide protection against people who use electronic systems to intimidate or harass another person. This could be limited to one-on-one interactions (individual perspective) or extend to the harassment of that person and others in their network (collective perspective).
17. *Extortion:* the aim is to make it punishable by law to extort a benefit from another person by threatening to publish computer data containing personal or private information which can cause the other person public ridicule, contempt, hatred or embarrassment.

## 4.3.2 The State of Enforced Law

### 4.3.2.1 Region Wide Review

The evolution of online safety legislation in the Caribbean has been slow and fractured [13, 53]. The results of the analysis of our corpus are summarized in figure 4.1 and table 4.2. Of the 13 countries included in the analysis, ten countries had substantive laws enforced, one had limited substantive laws enforced with a more comprehensive bill that has not been passed (after five years of being drafted), and two countries no dedicated laws enforced that specifically provided protections against online threats. A wide disparity was observed among the dates of enforcement. Earlier adopters like the Bahamas had laws enforced since 2003 while other like Belize enforced a comprehensive act in 2020. Thus, despite the adoption of model laws, there are wide differences in the levels of maturity in legislative approach and coverage. Moreover, only a limited number of countries (4/13) had established response teams for cyber-related offenses. The variation could be attributed to many socioeconomic factors such as GDP, population thus human capital, and delays in technology adoption. The countries that had dedicated response teams had on average populations of at least 100,000. Countries that did not have these teams in place were reliant on their local law enforcement agencies to address reports and proceeding the judicial process. This may be an appropriate allocation of resources for smaller populations such as St. Kitts-Nevis (<70,000). However, this approach may not be sustainable especially if there is not a national strategy towards online safety nor or are there consistent opportunities for adequate training or investment in technological equipment.

Unsurprisingly, countries with more recently passed pieces of legislation, on average had more categories of provisions. Furthermore, across the three major categories of offenses, computer-related offenses had that highest level of coverage. In contrast, there were serious variance in the coverage for interactional offenses. A full overview of the coverage of all provisions across all countries is provided in table 4.2. In the next section, we describe the extent of coverage per country.

### 4.3.2.2 Country Review

For each country, we wanted to provide a holistic review of the overall maturity of that country’s approach to online safety. Therefore, we employ the Cybersecurity Maturity Model (CMM) as an additional metric (see Rea-Guam et al. for overview on maturity models [149]). The assessment

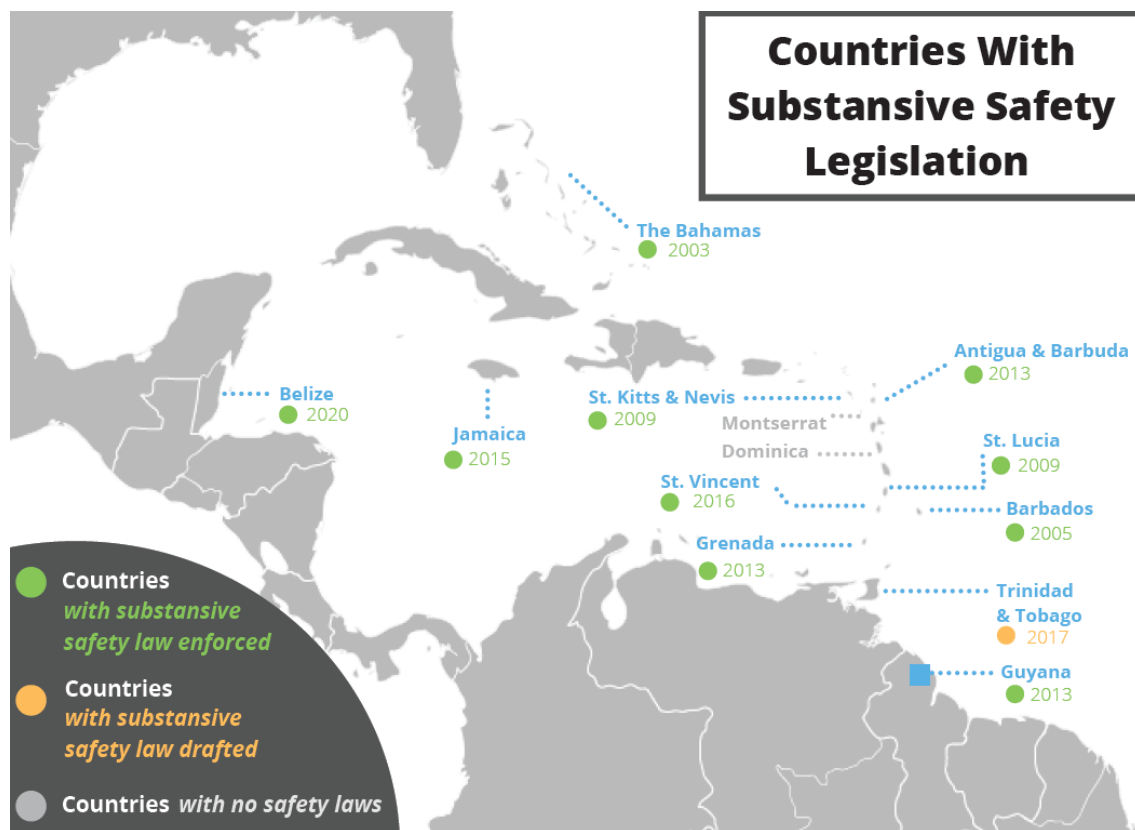


Figure 4.1: Overview of legislative protections related to the prevention and prohibition of online safety threats. Note: The figure displays "St. Vincent" which represents St. Vincent and the Grenadines.

Country	Law	Year	Status
Antigua & Barbuda	Electronic Crimes Act	2013	Enforced
	Data Protection Act	2013	Enforced
	The Computer Misuse Act	2006	Enforced
Bahamas	Computer Misuse Act	2003	Enforced
	Data Protection (Privacy of Personal Information)	2003	Enforced
	Electronic Communication and Transactions Act	2003	Enforced
	Sexual Offences and Domestic Violence Act	2010	Enforced
Barbados	Computer Misuse Act	2005	Enforced
	Barbados Data Protection Act	2019	Enforced
St. Lucia	Electronic Crimes Bill	2009	Enforced
	Data Protection Act	2011	Enforced
St. Kitts & Nevis	Electronic Crimes Act	2009	Enforced
	Data Protection Act	2018	Enforced
St. Vincent & the Grenadines	Cybercrimes Act	2016	Enforced
	Privacy Act	2003	Enforced
Trinidad & Tobago	Cybercrime Bill	2017	Drafted
	Computer Misuse Act	2000	Enforced
	Interception of Communication Act	2010	Enforced
Jamaica	Cybercrimes Act	2015	Enforced
	Data Protection Act	2020	Enforced
Dominica	None	-	None
Guyana	Cybercrimes Act	2018	Enforced
Grenada	Electronic Crimes Bill	2013	Enforced
	Interception of Communications Act	2013	Enforced
Belize	Cybercrime Act	2020	Enforced
	Belize Data Protection Act	2021	Enforced
	Telecommunication Act	2002	Enforced
	Electronic Transactions Act	2003	Enforced
	Interception of Communication Act	2010	Enforced
Montserrat	None	-	None

Table 4.1: Overview of regional legislative protections related to the prevention and prohibition of online safety threats and data protection

Category	Types of harms covered	St. Kitts & Nevis	Antigua & Barbuda	Bahamas	Barbados	St. Lucia	St. Vincent & the Grenadines	Trinidad & Tobago	Jamaica	Dominica	Guyana	Grenada	Belize
Computer-related offenses	Fraud	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓
Computer-related offenses	Illegal access to device	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓
Computer-related offenses	Illegal access to data	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓
Computer-related offenses	Data interference	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓
Computer-related offenses	Electronic Terrorism	✗	✓	✗	✗	✓	✓	✗	✗	✗	✗	✗	✗
Computer-related offenses	Spam	✓	✓	✓	✗	✓	✓	✗	✓	✗	✗	✗	✗
Content-related offenses	Privacy leaks/breaches	✓	✓	✗	✗	✓	✓	✓	✓	✗	✗	✓	✓
Content-related offenses	Pornography (revenge porn)	✗	✓	✓	✓	✓	✓	✓	✗	✗	✓	✓	✓
Content-related offenses	Pornography (child porn)	✓	✓	✓	✗	✓	✗	✓	✗	✗	✓	✓	✓
Content-related offenses	Identity theft	✓	✓	✗	✗	✗	✓	✓	✗	✗	✓	✓	✓
Interactional offenses	Online stalking	✗	✓	✗	✓	✓	✗	✓	✓	✗	✗	✓	✗
Interactional offenses	Online harassment (individual)	✗	✓	✗	✓	✓	✓	✓	✓	✗	✓	✓	✓
Interactional offenses	Online harassment (collective)	✗	✗	✗	✗	✗	✗	✗	✓	✗	✓	✗	partially
Interactional offenses	Defamation	✗	✓	✗	✓	✓	✓	✓	✓	✗	✗	✓	partially
Interactional offenses	Child luring	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗	✓
Interactional offenses	Extortion	✗	✗	✗	✗	✓	✗	✓	✗	✗	✗	✗	✗

Table 4.2: Regional Legislative Coverage regarding online offenses

tool is used to measure the level of maturity of a nation with regards to cybersecurity across five different dimensions (see 4.2 “Dimensions of the Cybersecurity Maturity Model”.) Each dimension provides several indicators of cyber capacity (an average of 10 indicators per dimension) in order for a nation to understand the stage of maturity in each specific consideration [46]. These indicators are measured across five levels of maturity: Start-up, Formative, Established, Strategic and Dynamic. The stages of maturity vary from an initial stage of maturity where a nation may have just begun to consider cybersecurity, through to a dynamic stage where a nation is able to quickly adapt to changes in the cybersecurity landscape, by balancing threat, vulnerability, risk, economic strategy or changing international needs, while at the same time improving its posture and readiness to face new threats. In the subsequent country profiles, the figure includes an indicator bar at the top of the figure that corresponds with the CMM rating. Additionally, information is provided about the key regulatory frameworks that define standalone offenses and any supporting statutes that offer online protection. We also report the agency that hosts the cyber response team (if there is one). The remainder of the figure describes the extent of coverage for the provisions considered across all countries. It should be noted that Montserrat was not assigned a profile since there was not enough evidence of concrete actions towards regulation.

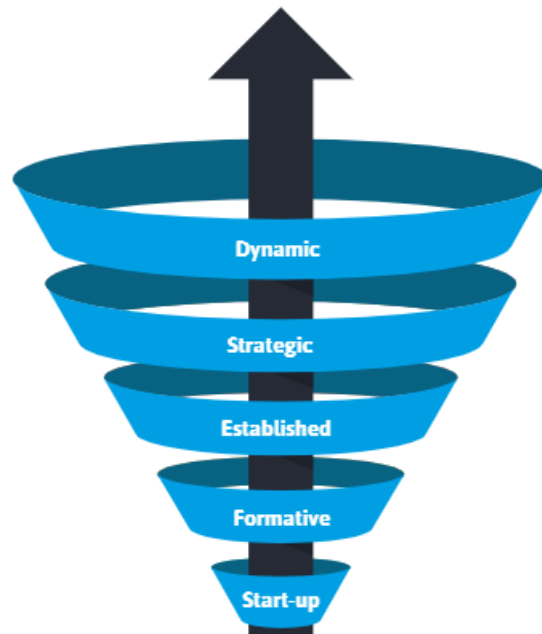


Figure 4.2: Dimensions of the Cybersecurity Capacity Maturity Model for Nations (CMM) from The Global Cyber Security Capacity Centre (GSCC). Source: Inter-American Development Bank <sup>1</sup>

**Antigua and Barbuda:** This twin island federation is located in the eastern Caribbean with a population of 97, 928<sup>2</sup>. Antigua and Barbuda has drafted and enforced legislation against online threats since 2013 with the Electronic Crimes Act of 2013 serving as the principle legislation defining protected provisions. In terms of coverage, this country offers protection across all computer-related and content-related offenses. The only exclusions are centered around a lack of explicit verbiage around extortion, child luring, and the harassment of persons within one's network (e.g. threatening family members). There is currently no cyber response team in place and as such local law enforcement is responsible for processing these offenses. As such, Antigua and Barbuda would be categorized within the formative stage based on CMM. This indicates although considerable progress has been made there is no set national strategy in place for responding to threats and improvements could be made on enforcement efforts. See figure 4.3 for this country's profile.

Antigua and Barbuda		
<b>KEY REGULATORY FRAMEWORKS:</b> Electronic Crimes Act (2013)		
<b>SUPPORTING STATUTES:</b> Data Protection Act (2013) The Computer Misuse Act (2006)		
Category	Offense	Coverage
Computer-related offenses	Fraud	✓
Computer-related offenses	Illegal access to device	✓
Computer-related offenses	Illegal access to data	✓
Computer-related offenses	Data interference	✓
Computer-related offenses	Electronic Terrorism	✓
Computer-related offenses	Spam	✓
Content-related offenses	Privacy leaks/breaches	✓
Content-related offenses	Pornography (revenge porn)	✓
Content-related offenses	Pornography (child porn)	✓
Content-related offenses	Identity theft	✓
Interactional offences	Online stalking	✓
Interactional offences	Online harassment (individual)	✓
Interactional offences	Online harassment (collective)	✗
Interactional offences	Defamation	✓
Interactional offences	Child luring	✗
Interactional offences	Extortion	✗

Figure 4.3: Overview of legislative protections related to the prevention and prohibition of online safety threats in Antigua and Barbuda.

**The Bahamas:** The Bahamas has been one of the earliest adopters to safety regulation. However, the country employs a more fractured approach to legal protection for online offenses. The

<sup>2</sup>Source: <https://datatopics.worldbank.org/world-development-indicators/>

Computer Misuse Act has a limited scope of definitions. As a result there is only one interactional offense that is covered. This could be challenging for persons wishing to pursue judicial options since precedent will be set by case law. Thus, specific verbiage or details of an unrelated case may contribute to additional barriers to ones case. There are protections under the penal code that offer protections for offenses such as revenge porn. The country has no dedicated cyber response team or an established nation cyber offense strategy. The Bahamas is categorized under the formative dimension as legislation is in place but improvement towards a more comprehensive strategy could be made. See 4.4 for this country’s profile.


 <b>The Bahamas</b>		
<b>KEY REGULATORY FRAMEWORKS:</b> The Computer Misuse Act (2006)		
<b>SUPPORTING STATUTES:</b> Data Protection (Privacy of Personal Information) (2003) Electronic Communication and Transactions Act (2003) Sexual Offences and Domestic Violence Act (2010)		
Category	Offense	Coverage
Computer-related offenses	Fraud	✓
Computer-related offenses	Illegal access to device	✓
Computer-related offenses	Illegal access to data	✓
Computer-related offenses	Data interference	✓
Computer-related offenses	Electronic Terrorism	✗
Computer-related offenses	Spam	✗
Content-related offenses	Privacy leaks/breaches	✓
Content-related offenses	Pornography (revenge porn)	✓
Content-related offenses	Pornography (child porn)	✓
Content-related offenses	Identity theft	✗
Interactional offences	Online stalking	✗
Interactional offences	Online harassment (individual)	✗
Interactional offences	Online harassment (collective)	✗
Interactional offences	Defamation	✓
Interactional offences	Child luring	✗
Interactional offences	Extortion	✗

Figure 4.4: Overview of legislative protections related to the prevention and prohibition of online safety threats in The Bahamas.

**Barbados:** Barbados is one of the few countries in the region with an established cyber response team responsible for the enforcement of online protections. The country has also made recent strides to develop data protection provisions with the Barbados Data Protection Act. There is not consistent coverage across any one category of harms. Barbados is categorized under the formative dimension as legislation is in place but improvement towards a more comprehensive strategy could be made. See 4.5 for this country’s profile.




 <b>Barbados</b>		
<b>KEY REGULATORY FRAMEWORKS:</b> Computer Misuse Act (2005)		
<b>SUPPORTING STATUTES:</b> Barbados Data Protection Act (2019)		
<b>CSIRT:</b> Barbados National Cyber Security Incident Response Centre (CIRT_BB)		
Category	Offense	Coverage
Computer-related offenses	Fraud	✓
Computer-related offenses	Illegal access to device	✓
Computer-related offenses	Illegal access to data	✓
Computer-related offenses	Data interference	✓
Computer-related offenses	Electronic Terrorism	✗
Computer-related offenses	Spam	✗
Content-related offenses	Privacy leaks/breaches	✓
Content-related offenses	Pornography (revenge porn)	✓
Content-related offenses	Pornography (child porn)	✗
Content-related offenses	Identity theft	✗
Interactional offences	Online stalking	✓
Interactional offences	Online harassment (individual)	✓
Interactional offences	Online harassment (collective)	✗
Interactional offences	Defamation	✓
Interactional offences	Child luring	✗
Interactional offences	Extortion	✗

Figure 4.5: Overview of legislative protections related to the prevention and prohibition of online safety threats in Barbados.

**Belize:** Belize has made considerable strides in recent years towards the development of the country's cybersecurity efforts. The country drafted and enforced the Cybercrime Act of 2020 and more recently enforced the Belize Data Protection Act of 2021. There are substantive provisions for all content-related offenses. However, the legislation does not explicitly define certain standalone offenses such as online stalking or have clear verbiage on defamation. The country has no dedicated cyber response team but they have an established nation cyber-offense strategy. Belize is the only country in this study categorized under the established dimension as legislation is in place and actions are being made about a more comprehensive strategy. See 4.6 for this country's profile.


 <b>Belize</b>		
<b>KEY REGULATORY FRAMEWORKS:</b> Cybercrime Act (2020)		
<b>SUPPORTING STATUTES:</b> Belize Data Protection Act (2021) Telecommunication Act (2002) Electronic Transactions Act (2003) Interception of Communication Act (2010)		
<b>CSIRT:</b> None		
Category	Offense	Coverage
Computer-related offenses	Fraud	✓
Computer-related offenses	Illegal access to device	✓
Computer-related offenses	Illegal access to data	✓
Computer-related offenses	Data interference	✓
Computer-related offenses	Electronic Terrorism	✗
Computer-related offenses	Spam	✗
Content-related offenses	Privacy leaks/breaches	✓
Content-related offenses	Pornography (revenge porn)	✓
Content-related offenses	Pornography (child porn)	✓
Content-related offenses	Identity theft	✓
Interactional offences	Online stalking	✗
Interactional offences	Online harassment (individual)	✓
Interactional offences	Online harassment (collective)	✗
Interactional offences	Defamation	partially
Interactional offences	Child luring	✓
Interactional offences	Extortion	✗

Figure 4.6: Overview of legislative protections related to the prevention and prohibition of online safety threats in Belize.

**Dominica:** This windward Caribbean country has a population of 71,991<sup>3</sup>. Similar to other neighboring countries, Dominica considered the Electronic Crimes Bill. This legislation was drafted in 2013 and as of 2022, it has not yet been passed. There is no cyber response team in place

<sup>3</sup>Source: <https://datatopics.worldbank.org/world-development-indicators/>

and no defined laws to offer protection. Dominica is categorized in the start-up dimension since regulatory planning is in its infancy stage. See 4.7 for this country's profile.


 <b>Dominica</b>		
<b>KEY REGULATORY FRAMEWORKS:</b> No major legislation enforced. The Electronic Crimes Bill of 2003 was drafted but has not been passed into law.		
<b>SUPPORTING STATUES:</b> NA		
<b>CSIRT:</b> None		
Category	Offense	Coverage
Computer-related offenses	Fraud	X
Computer-related offenses	Illegal access to device	X
Computer-related offenses	Illegal access to data	X
Computer-related offenses	Data interference	X
Computer-related offenses	Electronic Terrorism	X
Computer-related offenses	Spam	X
Content-related offenses	Privacy leaks/breaches	X
Content-related offenses	Pornography (revenge porn)	X
Content-related offenses	Pornography (child porn)	X
Content-related offenses	Identity theft	X
Interactional offences	Online stalking	X
Interactional offences	Online harassment (individual)	X
Interactional offences	Online harassment (collective)	X
Interactional offences	Defamation	X
Interactional offences	Child luring	X
Interactional offences	Extortion	X

Figure 4.7: Overview of legislative protections related to the prevention and prohibition of online safety threats in Dominica.

**Grenada:** Similar to neighboring countries, Grenada enforced their Electronic Crimes Act in 2013. Additional statutes that provide support include the Interception of Communications Act which includes protections for data in public and private contexts. Protections are offered across all content related offenses with some exclusions for coverage under interactional and computer related offenses. The country does not have a cyber response team or a nation online threat strategy. Therefore, it is categorized with the formative dimension. See the country's profile in figure 4.8.

**Guyana:** Guyana is country that has cultural ties to Caribbean countries geographically located within the larger Caribbean archipelago. However, Guyana is located in South America with a population of 786,559<sup>4</sup>. The country recently enforced their Cybercrimes Act of 2018. Unlike many countries included in this analysis, Guyana offers a wide coverage of interactional offenses. It is the

<sup>4</sup>Source: <https://datatopics.worldbank.org/world-development-indicators/>


 <b>Grenada</b>		
<b>KEY REGULATORY FRAMEWORKS:</b> Electronic Crimes Act (2013)		
<b>SUPPORTING STATUTES:</b> Interception of Communications Act (2013)		
<b>CSIRT:</b> None		
Category	Offense	Coverage
Computer-related offenses	Fraud	✓
Computer-related offenses	Illegal access to device	✓
Computer-related offenses	Illegal access to data	✓
Computer-related offenses	Data interference	✓
Computer-related offenses	Electronic Terrorism	✗
Computer-related offenses	Spam	✗
Content-related offenses	Privacy leaks/breaches	✓
Content-related offenses	Pornography (revenge porn)	✓
Content-related offenses	Pornography (child porn)	✓
Content-related offenses	Identity theft	✓
Interactional offences	Online stalking	✓
Interactional offences	Online harassment (individual)	✓
Interactional offences	Online harassment (collective)	✗
Interactional offences	Defamation	✓
Interactional offences	Child luring	✗
Interactional offences	Extortion	✗

Figure 4.8: Overview of legislative protections related to the prevention and prohibition of online safety threats in Grenada.

only country in our analysis that has provisions for the harassment of others within your network (e.g. friends and family). Additionally, it is one of the few countries that has explicit verbiage about defining extortion and child luring as a standalone offenses. Guyana has a cyber response team in place but no clear cyber protection strategy. Thus, it is categorized in the formative dimension. See 4.9 for Guyana's profile.

Guyana		
<b>KEY REGULATORY FRAMEWORKS:</b> Cybercrimes Act (2018)		
<b>SUPPORTING STATUTES:</b> NA		
<b>CSIRT:</b> Ministry of Public Security (CIRT.GY)		
Category	Offense	Coverage
Computer-related offenses	Fraud	✓
Computer-related offenses	Illegal access to device	✓
Computer-related offenses	Illegal access to data	✓
Computer-related offenses	Data interference	✓
Computer-related offenses	Electronic Terrorism	✗
Computer-related offenses	Spam	✗
Content-related offenses	Privacy leaks/breaches	✗
Content-related offenses	Pornography (revenge porn)	✓
Content-related offenses	Pornography (child porn)	✓
Content-related offenses	Identity theft	✓
Interactional offences	Online stalking	✗
Interactional offences	Online harassment (individual)	✓
Interactional offences	Online harassment (collective)	✓
Interactional offences	Defamation	✓
Interactional offences	Child luring	✓
Interactional offences	Extortion	✓

Figure 4.9: Overview of legislative protections related to the prevention and prohibition of online safety threats in Guyana.

**Jamaica:** This country has a population of 2.961 million <sup>5</sup>. Jamaica has the Cybercrimes Act of 2015 currently enforced to provide regulatory protections against threats that emerge in online spaces. The country also recently enforced the Data Protection Act of 2020 which provides provisions for the transmission of private data. In terms of coverage, the country has limited protections for content-related offenses. Recent advocates for the revision of protections to make a clearer path to justice for certain crimes such as revenge porn [14]. Jamaica has a dedicated cyber response team and also allows victims to submit reports online to ensure resources are used efficiently and to

<sup>5</sup>Source: <https://datatopics.worldbank.org/world-development-indicators/>

save time. This country is considered to be in the established dimension although there are areas improvement that could be made. See figure 4.10.


 <b>Jamaica</b>		
<b>KEY REGULATORY FRAMEWORKS:</b> Cybercrimes Act (2015)		
<b>SUPPORTING STATUTES:</b> Data Protection Act (2020)		
<b>CSIRT:</b> Ministry of Science, Energy, and Technology (JaCIRT)		
Category	Offense	Coverage
Computer-related offenses	Fraud	✓
Computer-related offenses	Illegal access to device	✓
Computer-related offenses	Illegal access to data	✓
Computer-related offenses	Data interference	✓
Computer-related offenses	Electronic Terrorism	✓
Computer-related offenses	Spam	✗
Content-related offenses	Privacy leaks/breaches	✓
Content-related offenses	Pornography (revenge porn)	✗
Content-related offenses	Pornography (child porn)	✗
Content-related offenses	Identity theft	✗
Interactional offences	Online stalking	✓
Interactional offences	Online harassment (individual)	✓
Interactional offences	Online harassment (collective)	✗
Interactional offences	Defamation	✓
Interactional offences	Child luring	✗
Interactional offences	Extortion	✗

Figure 4.10: Overview of legislative protections related to the prevention and prohibition of online safety threats in Jamaica.

**St. Lucia:** This country has a population of 183,629 and uses the Electronic Crimes Act of 2009 as the key regulatory framework for protection from online threats. More recently, the Data Protection Act of 2018 was enforced for additional support. St. Lucia offers a wide range of coverage across all types of offenses. The country does not have a cyber response team but has made efforts towards its development. This country is categorized as established. See 4.11.

**St. Kitts Nevis:** This twin island federation has a population of 53,192<sup>6</sup>. The country has the Electronic Crimes Act of 2013, revised in 2017, enforced to provide protections against online threats. Coverage-wise, St. Kitts and Nevis has a very limited range of defined standalone offenses for interactional offenses. Additional statutes from the penal code, Data Protection Act, and Computer Misuse Act provide support. However, in practise, case law has been used to establish

<sup>6</sup>Source: <https://datatopics.worldbank.org/world-development-indicators/>


 <b>St. Lucia</b>		
<b>KEY REGULATORY FRAMEWORKS:</b> Electronic Crimes Act (2009)		
<b>SUPPORTING STATUTES:</b> Data Protection Act (2018)		
<b>CSIRT:</b> None		
Category	Offense	Coverage
Computer-related offenses	Fraud	✓
Computer-related offenses	Illegal access to device	✓
Computer-related offenses	Illegal access to data	✓
Computer-related offenses	Data interference	✓
Computer-related offenses	Electronic Terrorism	✓
Computer-related offenses	Spam	✓
Content-related offenses	Privacy leaks/breaches	✓
Content-related offenses	Pornography (revenge porn)	✓
Content-related offenses	Pornography (child porn)	✓
Content-related offenses	Identity theft	✗
Interactional offences	Online stalking	✓
Interactional offences	Online harassment (individual)	✓
Interactional offences	Online harassment (collective)	✗
Interactional offences	Defamation	✓
Interactional offences	Child luring	✗
Interactional offences	Extortion	✓

Figure 4.11: Overview of legislative protections related to the prevention and prohibition of online safety threats in St. Lucia.

precedence for the pursuit of interactional offenses such as revenge porn. The country does not have an establish cyber response team. St. Kitts and Nevis is categorized as formative. See figure 4.12.


 <b>St. Kitts and Nevis</b>		
<b>KEY REGULATORY FRAMEWORKS:</b> Electronic Crimes Act (2013)		
<b>SUPPORTING STATUTES:</b> Data Protection Act (2013) The Computer Misuse Act (2006)		
<b>CSIRT:</b> None		
Category	Offense	Coverage
Computer-related offenses	Fraud	✓
Computer-related offenses	Illegal access to device	✓
Computer-related offenses	Illegal access to data	✓
Computer-related offenses	Data interference	✓
Computer-related offenses	Electronic Terrorism	✗
Computer-related offenses	Spam	✓
Content-related offenses	Privacy leaks/breaches	✓
Content-related offenses	Pornography (revenge porn)	✗
Content-related offenses	Pornography (child porn)	✓
Content-related offenses	Identity theft	✓
Interactional offences	Online stalking	✗
Interactional offences	Online harassment (individual)	✗
Interactional offences	Online harassment (collective)	✗
Interactional offences	Defamation	✗
Interactional offences	Child luring	✗
Interactional offences	Extortion	✗
Computer-related offenses	Fraud	✗

Figure 4.12: Overview of legislative protections related to the prevention and prohibition of online safety threats in St. Kitts and Nevis.

**St. Vincent:** This country has a population of 110,211 <sup>7</sup>. St. Vincent has the Cybercrimes Act of 2016 currently enforced to provide regulatory protections against threats that emerge in online spaces. The country also enforced the Privacy Act of 2003 which provides provisions for the transmission of private data. In terms of coverage, the country has limited protections for interactional offenses. St. Vincent and the Grenadines does not have a cyber response team. This country is considered to be in the established dimension although there are areas improvement that could be made. See figure 4.13.

**Trinidad and Tobago:** This Caribbean country has a population of 1.399 million people <sup>8</sup>. Trinidad and Tobago is dissimilar to other countries considered in this analysis since the key

<sup>7</sup>Source: <https://data.worldbank.org/country/vc>

<sup>8</sup>Source: <https://www.worldbank.org/en/country/trinidadandtobago>




 <b>St. Vincent &amp; The Grenadines</b>		
<b>KEY REGULATORY FRAMEWORKS:</b> Cybercrimes Act (2016)		
<b>SUPPORTING STATUTES:</b> Privacy Act (2003)		
<b>CSIRT:</b> None		
Category	Offense	Coverage
Computer-related offenses	Fraud	✓
Computer-related offenses	Illegal access to device	✓
Computer-related offenses	Illegal access to data	✓
Computer-related offenses	Data interference	✓
Computer-related offenses	Electronic Terrorism	✓
Computer-related offenses	Spam	✓
Content-related offenses	Privacy leaks/breaches	✓
Content-related offenses	Pornography (revenge porn)	✓
Content-related offenses	Pornography (child porn)	✗
Content-related offenses	Identity theft	✓
Interactional offences	Online stalking	✗
Interactional offences	Online harassment (individual)	✓
Interactional offences	Online harassment (collective)	✗
Interactional offences	Defamation	✓
Interactional offences	Child luring	✗
Interactional offences	Extortion	✗

Figure 4.13: Overview of legislative protections related to the prevention and prohibition of online safety threats in St. Vincent and the Grenadines.


 <b>Trinidad and Tobago</b>		
<b>KEY REGULATORY FRAMEWORKS:</b> Cybercrime Bill (2017, drafted not enforced)		
<b>SUPPORTING STATUTES:</b> Computer Misuse Act (2000) Interception of Communication Act (2010)		
<b>CSIRT:</b> Ministry of National Security (TTCsIRT)		
Category	Offense	Coverage
Computer-related offenses	Fraud	✓
Computer-related offenses	Illegal access to device	✓
Computer-related offenses	Illegal access to data	✓
Computer-related offenses	Data interference	✓
Computer-related offenses	Electronic Terrorism	✗
Computer-related offenses	Spam	✗
Content-related offenses	Privacy leaks/breaches	✓
Content-related offenses	Pornography (revenge porn)	✓
Content-related offenses	Pornography (child porn)	✓
Content-related offenses	Identity theft	✓
Interactional offences	Online stalking	✓
Interactional offences	Online harassment (individual)	✓
Interactional offences	Online harassment (collective)	✗
Interactional offences	Defamation	✓
Interactional offences	Child luring	✗
Interactional offences	Extortion	✓

Figure 4.14: Overview of legislative protections related to the prevention and prohibition of online safety threats in Trinidad and Tobago.

legislative framework was drafted since 2017 but has not passed yet. As such, the Computer Misuse Act . The country has a dedicated cyber response team although there is not a strong regulatory framework to help support their efforts. This country is categorized in the formative dimension since steps have been taken towards more work in the future. See 4.14 for this country's profile.

#### 4.3.2.3 Online Threats: Media Reporting

Across the region, there is currently no strategic commitment to reporting cyber-related prosecutions. Countries such as Belize have been able to publish crime statistics with the cybercrime division producing a report on crimes statistics. However, this effort is not consistent across the region. Therefore it is difficult to assess how often persons are persecuted under the protections. Therefore, we conducted a preliminary content analysis on media reports of online threats in the region with the goal of investigating the prevalence of reporting related online violations and the prosecution of these violations. Two researchers conducted a search of news articles online using the substantive provisions as the keywords. Articles had to be in English and included coverage of a country within the scope of the legal analysis. Both researchers had domain knowledge of Caribbean-based news sources to access credibility. In total, 122 news articles were included in the final corpus <sup>9</sup>. We categorize the articles based on the type of threat to assess the distribution of threats in reports. We present descriptive data that offer insights into the reporting distribution.

Overall, Guyana had the highest number reports across all countries which accounted for 16% of the articles (see figure 4.17. Of all the reports from Guyana, non-consensual explicit imagery was the focus for 60% of the reports. This trend was consistent throughout all of the countries (see figure 4.17. Non-consensual explicit imagery emerged as the most reported type of threat across English speaking CARICOM countries.

We also coded the type of threats based on the category of offense that we identified during the legal analysis - interactional, content-related, and computer-related offenses. The results show that the majority of reports were focused on interactional offenses with computer-related offenses accounting for the lowest number of reports (see figure 4.17 for the breakdown). This is in direct contrast to the legislative analysis which showed that on average countries defined clear standalone offenses for computer-related offenses but coverage for interactional offenses was checkered. We

---

<sup>9</sup>See full corpus: <https://docs.google.com/spreadsheets/d/14kXOKMH00lMuTotgMoQ71qmMI9xi-KxtcBfoquzuw2M/edit?usp=sharing>

### Count of Threat Type Among All Countries

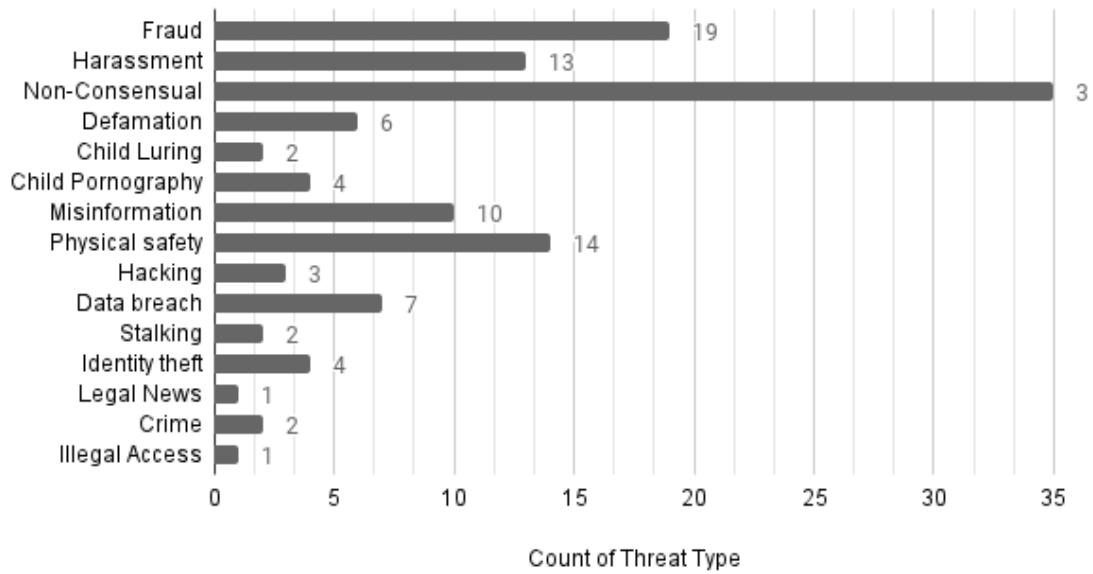


Figure 4.15: Results from the preliminary content analysis summarizing the total number of threats covered in the corpus organized by threat type

### Media Reporting of Online Threats by Country

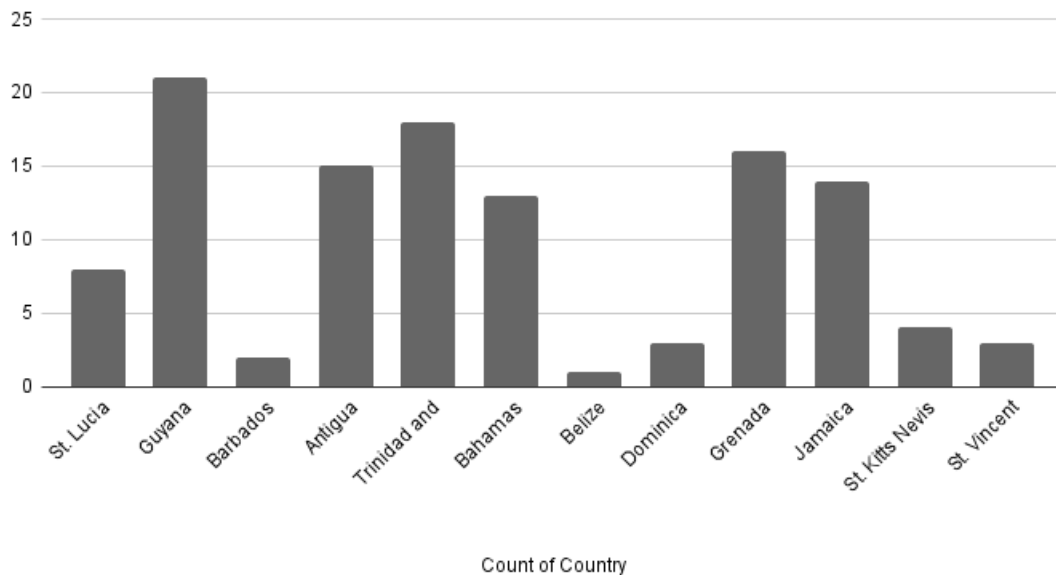


Figure 4.16: Results from the preliminary content analysis summarizing the total harms covered in corpus by country

discuss the implications of the findings and suggest pathways for future directions in the following section.

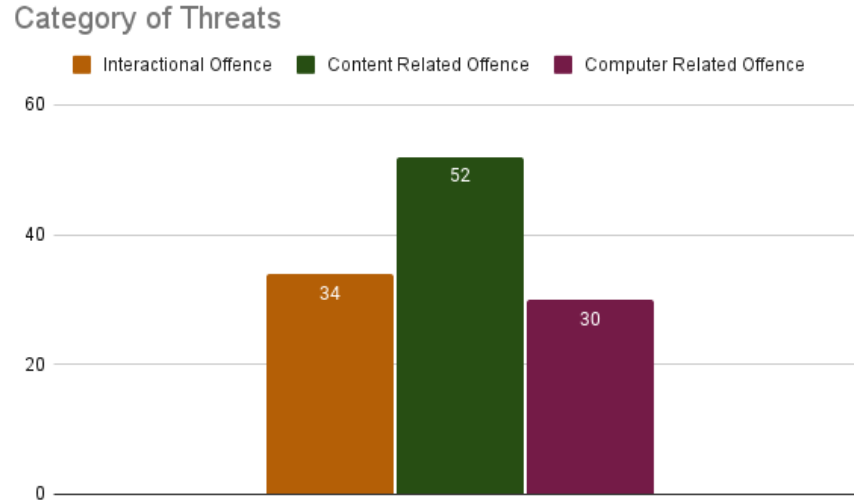


Figure 4.17: Results from the preliminary content analysis summarizing the total threats covered in corpus by threat category

## 4.4 Discussion

As summarised in Table 4.2, the coverage of standalone offenses varies significantly across CARICOM countries. Offenses such as illegal access to a device or data are explicitly defined as criminal across all member states. This trend is consistent among all computer-related offenses where the target is the data or device. There is a consistent trend of coverage regarding these types of offenses which is in stark contrast to offenses that cause physical and psychological harms such as harassment or cyberstalking. Arguably, legislators may opt for less explicit verbiage for flexibility in interpretation. This may prove to be beneficial as online threats continue to grown exponentially [86] and it may be difficult to prosecute offenders if the scope of the law is not representative of the harms people incur.

However, even if the offensive is not included, exclusion of the underlying harm itself may prove to be damaging. Countries that do not have explicit provisions for offenses rely on other statues which could produce imbalances in the perception of fairness as there may be limits and inconsistencies to the penalties of crimes that may be perceived as egregious. This is particularly

of concern for categories of offenses such non-consensual explicit imagery. We observed a consistent trend in media reports of victims of these offenses being women and young girls. Not explicitly having standalone offences that could be viewed as socially abhorrent could contribute to cycles of injustice for groups that are already vulnerable. In 2019, the Office of the Director of Public Prosecutions in Jamaica urged local lawmakers to specifically consider establishing a standalone offense for the distribution of non-consensual explicit imagery as current provisions under the Cybercrimes Act did not regard it as an offense [35]. Orrett Brown, deputy director of public prosecutions, stated that there were increasing incidents of this nature and that "this standalone offense that is being proposed will make it easier to prosecute that matter once we can establish that the image was published and the person did not consent to the image being published" [35].

Some level of variance in provisions is expected across member states. The criminalizing of specific types of offenses may be prioritized based on challenges that have emerged at a national level. Still, there is a need for a harmonized approach to behavior that might occur cross-borders. For example, only Belize and Guyana have provisions for protection against child luring. Yet, this is an offense that is not limited to the borders within the jurisdictions. Cybercrime officials have reprehended the growing number of non-Caribbean violators who have viewed the Caribbean as a hub for online violations against children [135]. Moreover, the legal protections offered assume the perpetrator would be human which limits provisions that could extend to biases and discrimination orchestrated by a system. This raises concern over the limits of prosecution for AI-enabled violations. Thus, establishing a regional legal instrument would be critical to respond to undesirable online behavior and outline an approach that would be representative of potential harms while affirming the rights of Caribbean citizens.

Beyond definitions, there are challenges that hinder the effective implementation of established laws. The lack of cyber response teams pushes the strain of investigation on to local law enforcement. This could be difficult for countries where there are also no dedicated departments for cybercrime prosecution. Thus, it would be of paramount importance to have law enforcement officers trained in evidence collection and sensitivity training to offer support for victims.

Below we summarize some of the key challenges that affect the effectiveness of a cohesive legislative approach to online safety and offer recommendations for addressing them:

- **Definitions and scope:** The omission of provisions that are common online threats could

cause friction for victims as they might be depending on case law or other statutes to establish a legal standing. **Recommendation:** Establish rights-affirming verbiage based on the harm that could be caused. This should cover a range not limited to the damage or loss of device but the impact that it could have to an individual and their network.

- **Enforcement and response:** The over-reliance on law enforcement officers in the region has been attributed to slower prosecution times, lower prosecution rates, and prosecutions falling to proceed due to "trans-jurisdictional barriers, subterfuge and the inability of key stakeholders in criminal justice systems to grasp fundamental aspects of technology aided crime" [14]. **Recommendation:** Establish clear investigative procedures and outline the major components needed. Building capacity for forensic tools would be critical. Stakeholders within CARICOM could consider resource sharing measures to ensure countries at different maturity levels could benefit from training. Moreover, this is an opportunity for technology companies to build better supporting tools that could assist in data preservation.
- **Access and Visibility:** All of the laws reviewed in our analysis were publicly available. However, for the average citizen, it may be difficult to know where to start, how to interpret provisions, and what rights they have. Persons may also not have the financial means to seek legal representation. This is a key consideration for the region considering the average income of Caribbean workers (see previous work in Chapter 2). **Recommendation:** Create more visibility of the rights afforded by each country. This could be done by combining the efforts of non-governmental organizations already making considerable strides towards online safety in the Caribbean. Get Safe Online<sup>10</sup> has a web presence that offers recommendations for protection online with ambassadors in all CARICOM member states. It would be helpful to utilize the existing network to create more public awareness on the rights that are in place to protect them. Additionally, cost-effective strategies for reporting would ease the burden on law enforcement and victims. For example, Jamaica has adopted technological reporting mechanisms for cybercrimes which eliminates the need for persons to physically journey to make a police report. Employing these methods could also assist in having more transparency in reports of cybercrimes as there is currently no way to substantiate the number of reports or prosecutions under the provisions.

---

<sup>10</sup><https://www.getsafeonline.org/>

## 4.5 Chapter Conclusion

In this chapter, we conducted a comparative legal analysis of crimes related to online safety throughout CARICOM member states to better understand the extent of legislative protections available to Caribbean citizen. We found that the majority of countries had some form of legislation in place. However, the extent of protections were often focused on threats faced in the Web 1.0 era and lack consistent protections for interpersonal threats. We outline areas for priority and offer recommendations geared at stakeholders who may be legislators, law enforcement, developers and designers. In the next chapter, I consider the gaps in the enforcement of available to consider how design could be used to enhance the process of pursuing justice. I apply the insights gained from this chapter and chapter 3 and propose to investigate affordances that contribute to more equitable options for protection online.



## Chapter 5

# Investigating the Role of Fairness, Equity, and Trust in Justice-Oriented Safety Interventions

Across the globe, there has been increased public pressure for social media platforms to take action against rising levels of online threats [86, 136, 150, 173, 198]. As illustrated in chapter 3, the Caribbean has not been immune to this pervasive phenomenon. Moreover, the supporting evidence from chapter 4 point towards opportunities to design alternative options for justice that depart from solely punitive approaches. Therefore, in this chapter I combine the insights from my previous work to explore how justice theories can inform how social media companies and communities respond to online offenses.

### 5.1 Background

Currently, social media platforms rely on several layers of safety mechanisms, moderators, and policy teams that actively work to develop an evolving set of platform-wide rules to detect

violations. Scholars have identified that these violations vary both in type and in severity [87, 86]. Various machine learning techniques have been developed and implemented to help reduce the risks to online safety in social media [145, 108, 100, 116, 64]. Platforms have implemented both proactive and reactive approaches. Reactive systems are triggered when a user already identifies problematic content which is brought to the platform’s attention, and it is then evaluated based on the policies and standards of that platform. The effectiveness of these methods have been criticized since the success of the technique is reliant on users flagging content. This may increase the possibility of risky content being circulated before finally being flagged. With reactive approaches, options for justice could include punitive methods such as removing content or banning users [158].

On the other hand, proactive approaches can include both manual and automated methods. These may include delaying the publication of content until they are evaluated by a human, the use of filters that prevent potentially problematic content from being posted, evaluating posting behavior to proactively block spam, or network-level signals such as IP addresses [64, 129, 130]. Proactive techniques have been used, for example, in the detection of potentially illegal objects in images and to reduce the intentional distribution of unsolicited images [73]. Moreover, AI techniques have been deployed to automate content moderation with the goal of reducing other malicious activities such as the prevalence of fake news (see [45] for an overview). One of the major challenges with AI-supported safety countermeasures is that a successful deployment is dependent on big but diverse data sets. Regardless, it is important to study how countermeasures (whether they be AI-supported or not) can be applied in a manner that addresses harms in an equitable but just manner.

Currently, social media companies place their efforts primarily around punitive justice where violators are silenced and expelled to achieve compliance with safety guidelines without addressing the underlying causes of the harm [130]. Hasinoff et al. argue that “the problem with these approaches to harm in online spaces are similar to the well-known limitations of the criminal legal system. Punishment itself is generally ineffective as a deterrent for those who harm others and rarely addresses the needs of those who have been harmed. This punitive system also does not encourage offenders to learn about the harm they have done and work to repair it, nor does it change the conditions and norms that facilitated the harm in the first place” [72]. Thus, punitive approaches to online safety could have immediate benefits to users but the downsides open vulnerabilities via increased exposure to violations and a disregard for users’ needs for recovery and healing. Unless there are approaches that provide alternatives, the response to online harms would continue echoing

Western criminal justice systems where the main perspective is removing offensive content and those who distribute them.

Rather than solely punitive approaches, in this chapter I argue for a different direction based on the principles of distributive and transformative justice. Transformative justice is regarded as a philosophical strategy to respond to violations by regarding the incident as an opportunity to address the root causes of conflict through growth and development. The focus is placed on interests rather than entitlements and claims [143]. In a different regard, distributive justice theories "give an answer to the question of how a society or a group should allocate its resources among individuals with competing needs or claims" [185]. By building on these principles, I present justice-oriented countermeasures for online safety. The designs offer an opportunity to address systemic challenges within the judicial system while acknowledging that victims often are better served by systems that meet their safety goals. In evaluating the effectiveness of these designs, this chapter seeks to answer the following research questions:

**RQ1:** How do different justice-oriented countermeasures influence users' perceived safety within online communities?

**RQ2:** How does personalization affect users' perceptions of justice-oriented countermeasures?

**RQ3:** How do justice-oriented countermeasures influence the adoption of protective behaviors within online communities?

In the following section, I offer a discussion of the theoretical motivations that inform the hypotheses tested in the study.

## 5.2 Hypotheses Development

### 5.2.1 Evaluating System Fairness

Multiple studies within the AI domain have highlighted the importance of evaluating system fairness as a key principle in ethical development [175, 138, 177, 88, 59, 120, 137]. Jobin et al. presented an analysis of major AI ethical guidelines and places justice and fairness as a prominent principle that is critical in the prevention, monitoring or mitigation of unwanted bias, discrimination, and reducing the proliferation of abuse [88]. However, the concept of fairness is continuously evolving to maintain more robust designs and understandings of justice. Social computing researchers have

highlighted the fluidity of fairness and justice [51, 42, 153]. Bennett and Keyes argue that adopting "a singular idea of "fairness" risks reinforcing existing power dynamics" through gatekeeping or the promotion of tools that harm those who are particularly vulnerable [20]. Systems could systematically reinforce inequitable distributions of resources which would place a burden on people who are already marginalized [79]. Consequently, design concepts around *distributive fairness* promote a "re-distribution of the production mechanisms for technology and information" [51]. This includes the departure from universalist designs and towards a direction where someone's context is considered to best deliver the appropriate resources needed [42]. People are more likely to perceive outcomes as being fair if they believe the system allocates the resources to allow them to effectively do that [65]. Likewise, *transformative fairness* encapsulates ideas around accountability and transformation for people who abuse and do harm through community healing [185]. Designing for transformation allows both parties to be involved in opportunities for growth and thus more robust and fair options for justice and sense-making when things go wrong [199]. *Procedural fairness* is closely related to the belief that fair procedures result in acceptable outcomes and as a result they are more willing to interact with products even if the results are unfair [168]. Therefore, the proposed study posits that the presence of justice-oriented countermeasures would influence users' perceptions regarding the evaluation of the system's fairness. Thus, suggesting supporting evidence towards possible recovery actions is expected to positively influence users' evaluation of the system's fairness. In a similar sense, providing personalized experiences have been shown to increase satisfaction and over user experience during decision-making [111]. As such, personalized countermeasures should be perceived more positively regarding how resources are allocated and how procedures are provided. I hypothesize:

**H1a:** Justice-oriented countermeasures will directly influence the system's fairness evaluation.

**H1b:** Justice-oriented countermeasures will directly influence perceptions of equity.

**H1c:** Personalization will positively influence the system's fairness evaluation.

### 5.2.2 The Role of Trust, Equity, and Control

Pillai et al. argues that higher levels of trust are upheld when distributive outcomes are perceived to be fair. In other words, people build trust in systems once they derive value around how well resources are allocated to arrive at fair outcomes. In a similar sense, prior work has

also shown support for the relationship between procedural fairness and trust. Certain rules and processes related to how platform violations are defined and subsequently remedied would influence users' perceptions of the level of fairness offered during the recovery process after a violation. As shown in chapter 3, once violations occur, there could be severe consequences to users' physical, emotional, digital safety. As such, increasing perceptions of fairness could enhance users' trust and subsequently their attitudes towards protective behaviors. Additionally, procedural fairness has been shown to significantly influence trust development in online systems [68]. Unless people feel like they understand the procedures that influence outcomes of a system, they are less likely to consider the system as fair and would therefore require more control [50]. Likewise, transformative principles place emphasis on outcomes that reflect the interests of users by addressing systemic issues at the root [47]. By rendering options for growth, development, and healing, I anticipate an association with increased levels of benevolence. Therefore, I propose:

**H2a:** Fairness evaluation (distributive, procedural, and transformative fairness) will be positively associated with increased levels of trust

**H2b:** Fairness evaluation (distributive, procedural, and transformative fairness) will be positively associated with increased levels of perceived control

The equity theory posits that persons who are treated fairly by a party and display satisfaction with their experience or outcome are more likely to engage in repeated behavior [4]. In line with these findings, I hypothesize the following:

**H2c:** Fairness evaluation (distributive, procedural, and transformative fairness) will be positively associated with perceived equity

### 5.2.3 Effects on Intention to Adopt

Design approaches that address systemic imbalances and acknowledge users' unique contexts are more likely to be perceived as more engaging, satisfying, personalized experiences [42]. As an individual experiences stronger attitudes towards how well the system working in supporting their post-violation recovery and maintaining safety, they will be more motivated to engage in protective behavior [118]. In a similar way, a user who is more confident in their ability to effectively use a tool is more likely to be positively motivated to engage with that tool [122]. Therefore:

**H3a:** Perceived control will positively be associated with behavioral intention.

**H3b:** Trust will positively be associated with behavioral intention.

**H3c:** Perceived equity will positively be associated with behavioral intention.

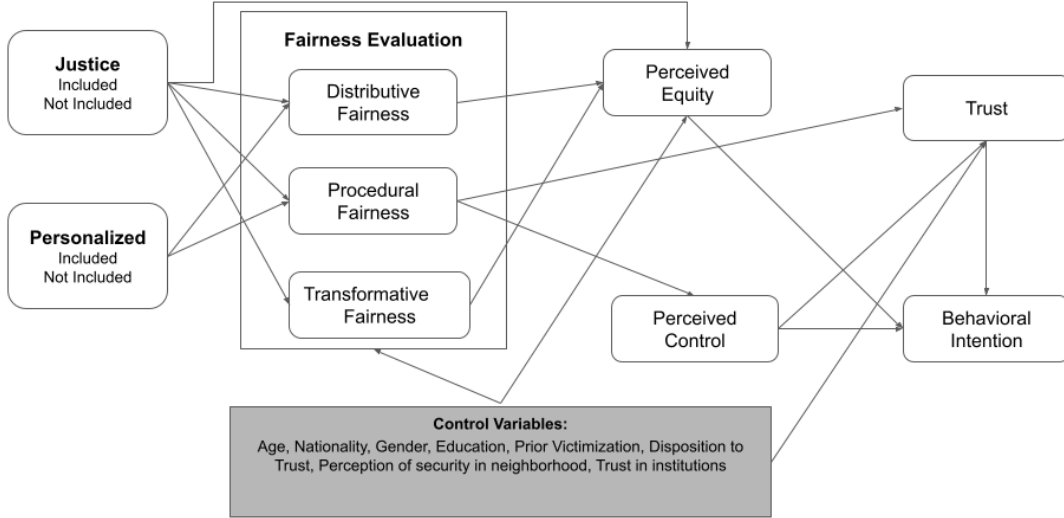


Figure 5.1: Proposed conceptual model for the study

## 5.3 Method

### 5.3.1 Overview

In this study, an online experimental user study was conducted using a self-developed prototype of a social media site called *Community*. The following sections detail the study procedure as well as details about recruitment.

### 5.3.2 Stimuli, Design and Procedures

The stimuli for the study consisted of two parts: a social media post with offensive content that violates *Community* standards and a safety countermeasure to remedy the violation. A total of five posts were shown to participants with varying levels of severity for each of the violations. To gauge the level of severity of the various harmful content, I categorize harms based on Caribbean citizens' self-reported levels of different harms in Chapter 3 and cross-reference international perceptions of severity based on work by Jiang et al [87].

To increase realism, the layout and key features of the *Community* prototype was similar to existing social media sites. Users were shown violations of *Community* standards and asked to interact with safety countermeasures (see example scenarios in Figure 5.3). Also, scenarios had details that were specific to the Caribbean context. For example, one scenario mentioned having a bank account hacked but included the name and logo of a bank that operates regionally and is present in all of the countries where data was collected (see Figure 5.3b). The overall procedures for the study are illustrated in Figure 5.2.

After completing training, participants were randomly assigned to one of the experimental conditions where they were presented with five (5) scenarios with violations and asked to interact with the corresponding safety countermeasure. All scenarios are included. Reaction statements related to the perception of countermeasure was presented with each scenario to gather subjective data on the countermeasure. The order of the violations are randomized to reduce potential order effects. After all scenarios were presented, participants were asked to complete a post-stimulus survey. Section 5.3.5 describes the measurement items in detail.

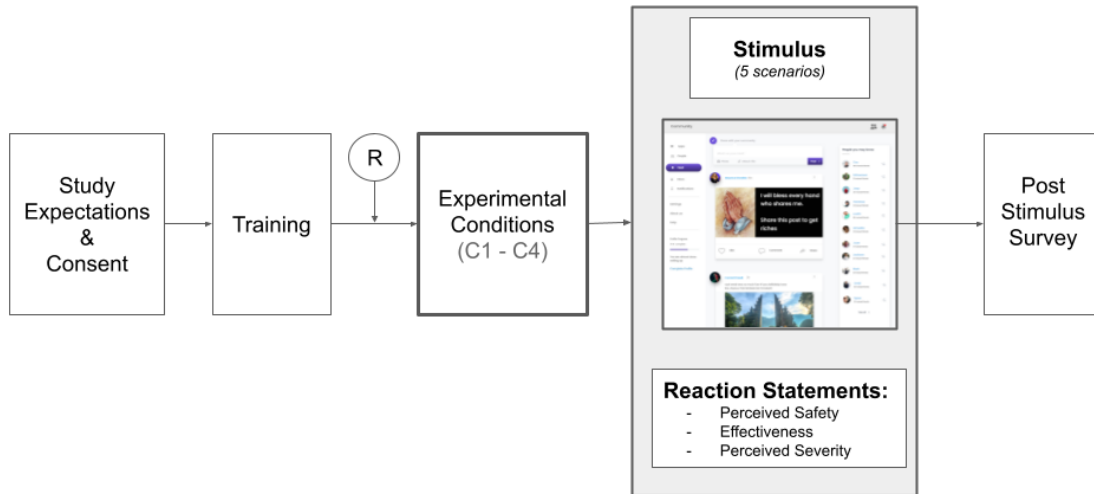


Figure 5.2: An overview of the study procedure. "R" denotes that participants will be randomly assigned to one of the four conditions.

### 5.3.3 Experimental Manipulations

This section describes the independent variables that were employed in the proposed study. The experiment followed a 2 x 2 between-subjects design where justice was manipulated on two levels (included or not included) and personalization was also manipulated on two binary levels (personalized or not personalized).

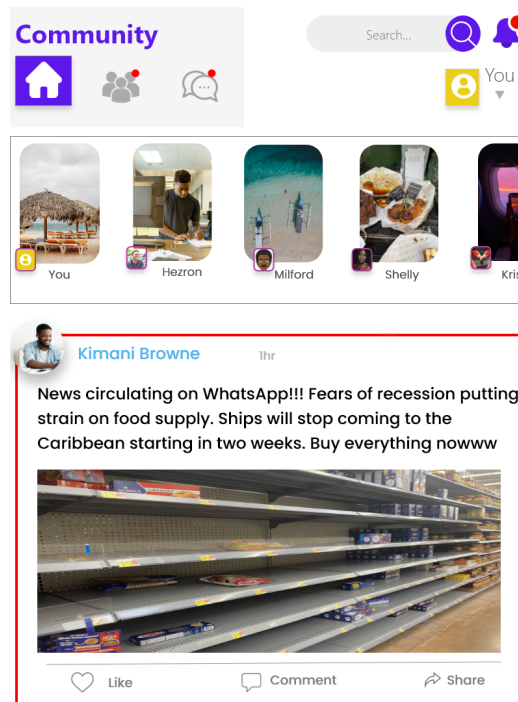
#### 5.3.3.1 Independent Variable: Justice-Oriented Countermeasure

Within the contexts of this study, *justice* was designed with core principles from existing criminal justice theories. Specifically, we consider principles informed by transformative and distributive justice that emphasize accountability, actionability, and equitable resource allocation. Thus, an alternative justice-oriented countermeasure would "identify what harm has been done; who is involved and impacted; what their resulting needs are; and what future actions are needed to heal the traumas resulting from an act of harm and address the needs of those affected" [12, 143]. We manipulate these design principles by varying the presence of information related to accountability (harm identification), actionability (outlining options for redress), and equity (offering resources that acknowledge one's context). Figure 5.4, provides an illustration of a design that is justice-oriented compared to one that is not (figure 5.5). Both of these designs are not personalized to further highlight the differences among conditions.

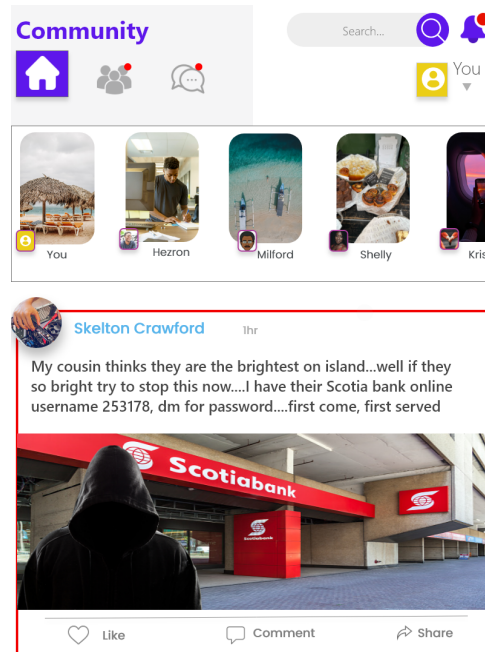
#### 5.3.3.2 Independent Variable: Personalization

Prior work has provided supporting evidence for the positive effect of personalized interfaces during the decision-making process [111]. In the designs, I consider personal context, specifically, geo-location to personalize countermeasures. Personalized designs for each respective country has detailed information related to that particular country. For example, participants from St. Lucia has legal information that was gathered from chapter 4 about fines and penalties related to the specific violation being presented. The information is adapted for each respective country based on current protections available. I manipulate personalization as a binary variable where insights about the options for redress are detailed. Figure 5.6 shows an example of a personalized design.





(a) Example scenario around misinformation



(b) Example scenario around leaked banking credentials

Figure 5.3: Example scenarios presented in the study

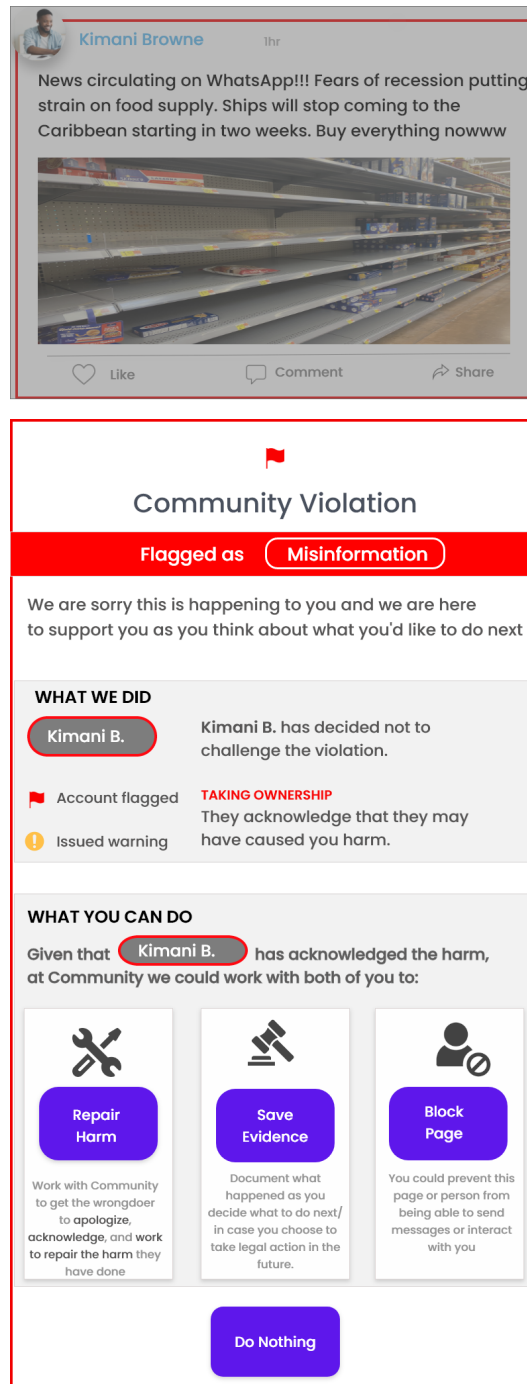


Figure 5.4: Example non-personalized design with a justice-oriented countermeasure

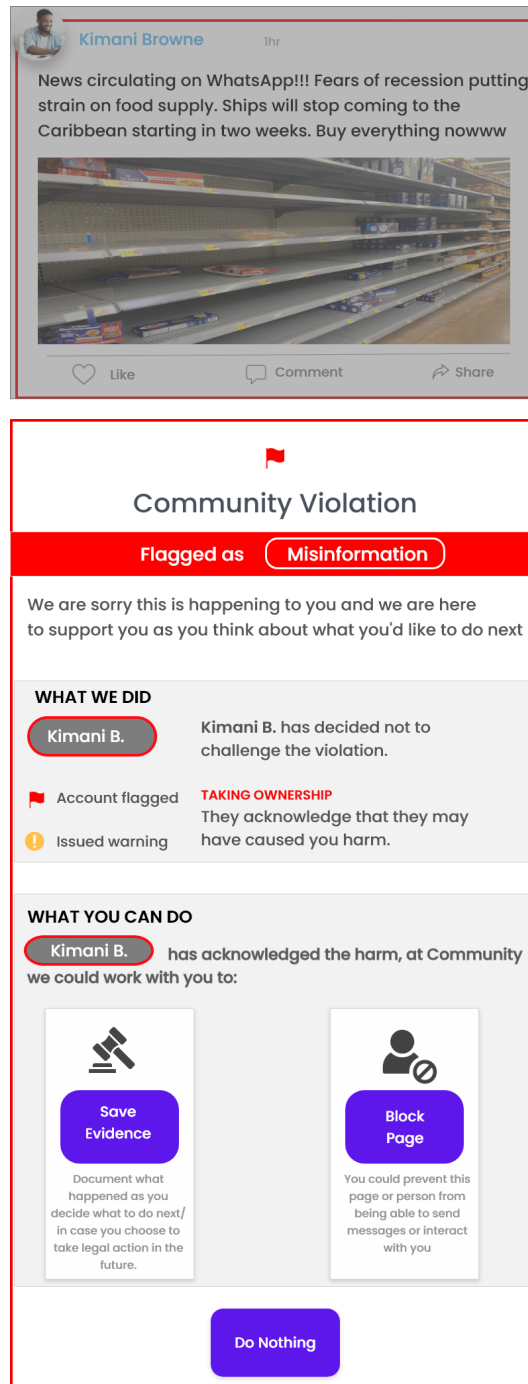
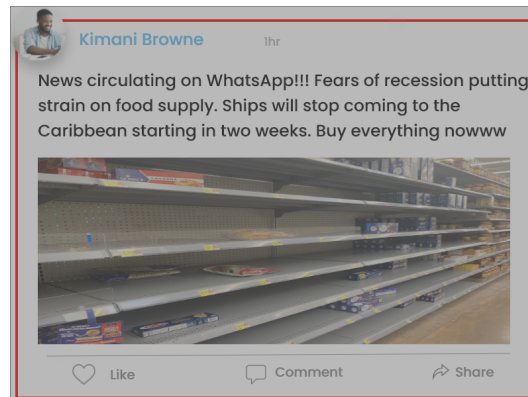



Figure 5.5: Example non-personalized design where an alternative justice countermeasure is not included







### Community Violation


Flagged as **Misinformation**

We are sorry this is happening to you and we are here to support you as you think about what you'd like to do next

#### WHAT WE DID


 **Kimani B.** Kimani B. has decided not to challenge the violation.

 Account flagged **TAKING OWNERSHIP**  
They acknowledge that they may have caused you harm.

 Issued warning


#### WHAT YOU CAN DO

Given that **Kimani B.** has acknowledged the harm, at Community we could work with both of you to:




**Repair Harm**

Work with Community to get the wrongdoer to apologize, acknowledge, and work to repair the harm they have done



**Save Evidence**

In **Jamaica**, people could be prosecuted for **misinformation** and be fined up to **\$1,500,000 JAM** or face **three months in prison**.



**Block Page**

You could prevent this page or person from being able to send messages or interact with you

**Do Nothing**

Figure 5.6: Example personalized justice-oriented countermeasure

### 5.3.4 Recruitment

One of the goals of the study was to identify country-level differences in users' perceptions of the justice-oriented designs. As such, we focus on countries with varied socioeconomic and legislative standings as countries are not homogeneous, underscoring the need to compare and assess the robustness of the conceptual model across context. From Chapter 3, we learned that there were country-to-country differences in willingness to adopt protective behaviors online. Countries such as St. Lucia exhibited significantly higher levels of willingness to adopt protective behaviors while also having increased levels of trust in platforms. This was in stark contrast with countries such as Jamaica which had significantly lower levels of trust and lower levels in intention to adopt protective behaviors.

Studying different national contexts is relevant because approaches on online safety differ substantially in law throughout the region. Based on the legislative analysis conducted in chapter 4, these two countries differ in regards to the extent of judicial protections offered to its citizens. St. Lucia has a stronger approach to online safety protection while Jamaica is in the process of making laws enforceable. I hypothesize that socioeconomic indicators and individual differences such as nationality will influence participants' attitudes towards the countermeasures. Thus, respondents were recruited from Caribbean countries with varying socioeconomic backgrounds and differing socio-technical approaches to online safety. A marketing research firm was hired to recruit participants from four Caribbean countries: Jamaica, Guyana, St. Lucia, and St. Kitts-Nevis. Remuneration was offered in local currencies but rounded closely to average \$5USD for completion of the study.

#### 5.3.4.1 Sample Description

A power analysis was conducted by using G\*Power <sup>1</sup> (See Figure 5.7 for result). The lower-bound for the sample size of the study was calculated using the following parameters: probability level set to  $\alpha=.05$ , desired statistical power level to 0.8, and a medium effect size ( $f=0.25$ ). The minimum sample size required to detect a medium effect would be 128 observations and 787 observations for small effect size ( $f=0.10$ ). Considering potential failures in the attention checks, I accounted for an additional 15% increase in sample size which brought the quota to 588 across all four countries. Quality checks were done based on duration to complete the study, low quality qual-

---

<sup>1</sup>G\*Power is a statistical program used to conduct power analyses. See: <https://www.psychologie.hhu.de/arbeitsgruppen/allgemeine-psychologie-und-arbeitspsychologie/gpower.html>

itative results, or evidence of straight lining. After data cleaning, the total sample size was 525: 155 from Jamaica, 136 from Guyana, 135 from St. Lucia, and 99 from St. Kitts-Nevis. Among the 525 participants, 64.3% were female, with a mean age of 28 years old (SD= 11.26; Median= 31 years old; Range= 18–67 years old), with 57% reported having completed post-high school education. Most participants identified as Black (63.6%), followed by Mixed races (10%), and East Indian (6%).

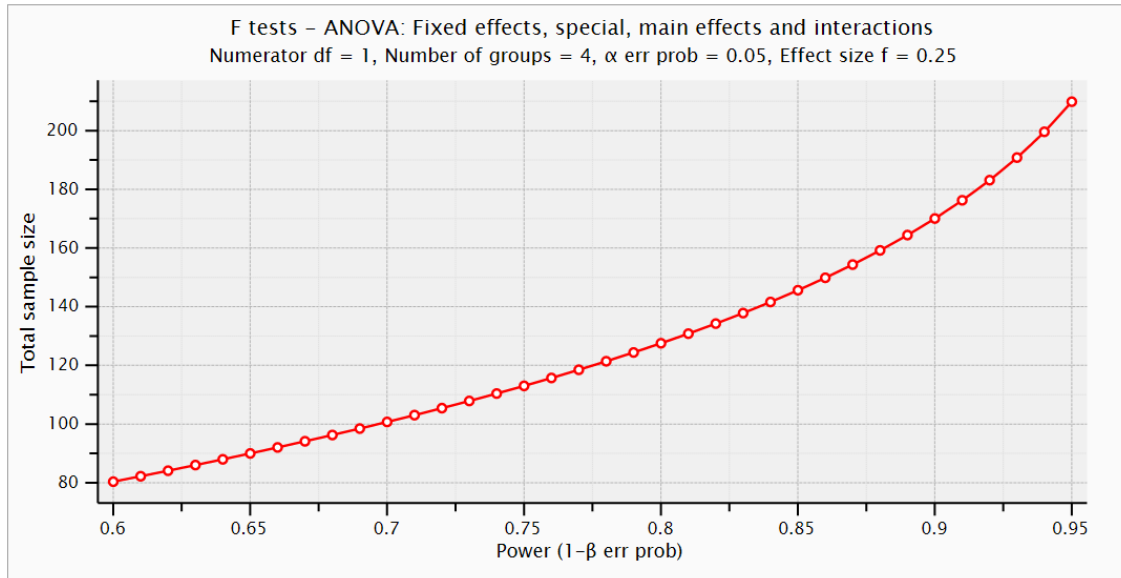


Figure 5.7: Sample size of the proposed study with a moderate effect size

### 5.3.5 Measurements

After being presented with the stimulus, participants were directed to complete the post-stimulus survey. The items were presented as seven-point likert scales. All items were adapted from previously validated scales. I utilized confirmatory factor analysis (CFA) to validate the structure of the model. The model was then subject to structural equation modeling (SEM). Applying CFA and SEM as an analytical method is beneficial in testing relationships between latent variables and validating conceptual models. Thus, the following constructs were considered for the model:

#### 5.3.5.1 Distributive Fairness

Distributive fairness is focused on how a society or a group should allocate its resources among individuals with competing needs or claims [99]. Hence, this construct assesses the extent to

which participants perceive the countermeasures or outcome to be proportionate and considerate of the needs of the parties. Five items were adapted from Verdonschot et al and Smith et al [185, 166]. Items include "Overall, the outcomes I received from the system were fair" (see all items included in this construct in Table 5.1).

#### **5.3.5.2 Procedural Fairness**

Procedural fairness is related with how users evaluate the recovery process after a violation by making judgments about the process of pursuing justice (i.e., the procedures, policies, and methods used by the system to address a problem). Exemplar items include "The social media platform gave me an opportunity to have a say in the handling of the problem." Six items were adapted from Sohaib et al and Grégoire and Fisher [168, 68].

#### **5.3.5.3 Transformative Fairness**

To measure transformative fairness, I adopt two items from Verdonschot et al [185] and introduce four new items. Transformative fairness is centered on fundamental transformation of the relationship between disputants by focusing on the underlying problem through growth and development. Items included: "The outcome would improve the damaged relationship with the other party that caused the problem".

#### **5.3.5.4 Perceived Equity**

The perceived equity scale was adopted from Colquitt, Verdonschot et al and Haynes et al [41, 185, 74]. Items measured the extent to which the system interventions acknowledged different resources needed from persons with differing needs. For example: "The social media platform provides interventions that acknowledge my culture, ethnicity, and identity".

#### **5.3.5.5 Perceived Control**

The perceived control scale measures the extent to which people believe the system influences their ability to exert control over situations or events. Items were adapted from Lee and Benbasat [106]. Example: "I think I had control over the outcome in the intervention process".

#### **5.3.5.6 Trust**

The trust scale measures the extent to which participants believe the system has integrity and works in their best interest. Items were adapted from [102]. Items include "I believe the social media platform would be open and receptive to the needs of its members."

#### **5.3.5.7 Behavioral Intention**

Behavioral intention has been shown to be a reliable metric for predicting actual behavior [184]. Thus, this construct measures participants' willingness to adopt justice-oriented mechanisms. Six items were adapted from [7, 49, 78].

#### **5.3.5.8 Control and Moderating Variables**

In this study, multiple variables were considered for a variety of causes that could explain possible variance of fairness related evaluations, perceived equity, trust, perceived control and behavioral intention. First, I controlled for the effects of age, gender and education, following HCI recommended guidelines for asking about identities [170]. Second, I considered the effects of the existing attitudes and beliefs around trust in institutions that should offer protection (such as judicial systems, law enforcement, social media platforms etc) as well as people's general disposition to trust [97]. I also controlled for perceptions of security in people's personal neighborhood as well as their prior victimised experiences [117]. Lastly, I consider potential country-to-country differences and observe nationality as a potential moderating variable.

## **5.4 Results**

This section presents the results of the online experiment. Data was triangulated based on: the reaction statements after interaction with each scenario, the measurement model, and qualitative data from free responses. For conciseness and clarity, I organise the results based on the research questions established in section 5.1.



### 5.4.1 RQ1: Do different justice-oriented countermeasures influence users' perceived safety within online communities?

To evaluate overall perceptions of safety, participants were asked about the how safe the countermeasure made them feel for each of the five scenarios presented. The results reveal that participants who had personalized justice-oriented measures (C1) rated feeling more safe while using the system compared to all other conditions. In a similar sense, participants with non-personalized countermeasures without a justice-oriented focus (C4) had the lowest scores overall regarding safety perceptions. Figure 5.8 illustrates the variance in safety scores across each condition. There were similar trends in safety perceptions for the different scenarios (see Figure 5.9).

Figure 5.8: Illustration of the variance in sum score for safety perceptions across all four conditions

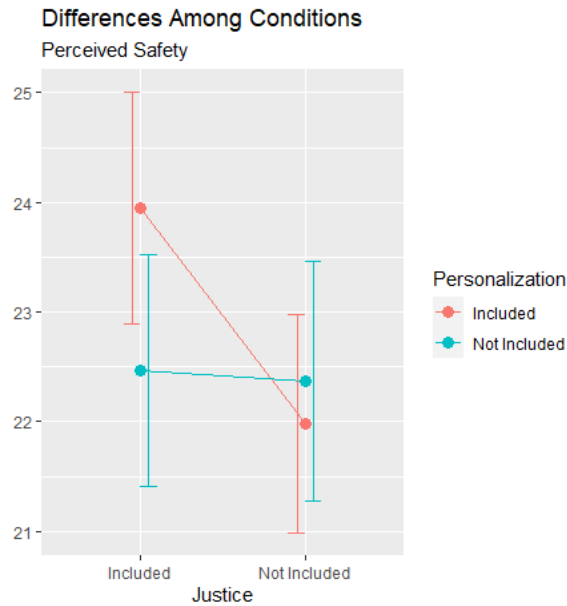
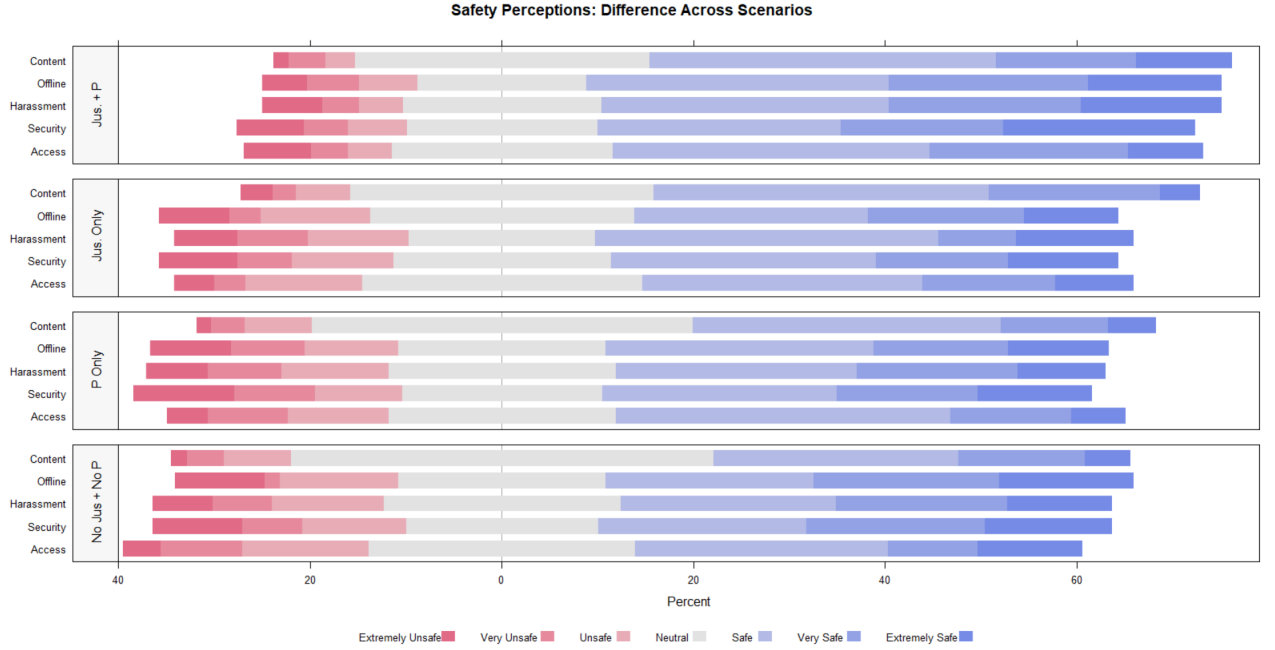


Figure 5.9: Variance in safety perceptions across all scenarios



#### 5.4.2 RQ2: Do justice-oriented countermeasures affect users' perceptions of fairness especially when personalized?

Investigating the effect of the manipulations required validating the robustness and validity of our measurement scales. Therefore, Confirmatory Factor Analysis (CFA) was employed using the lavaan package in R <sup>2</sup>. The results of the CFA are reported in Table 5.1. Items with low loadings were removed from subsequent analyses. Discriminant validity was confirmed and assessed by comparing the average variance extracted (AVE) of each factor against its correlation with other factors. All factors were evaluated for high reliability and convergent validity: Cronbach's  $\alpha$  values were excellent<sup>3</sup>, ranging between .85 and .92 while all AVE values exceeded 0.50. Once the CFA was completed, I applied structural equation modeling (SEM) to test the relationships between factors, as hypothesized by theory. SEM combines confirmatory factor analysis and path analysis to test hypothesized causal relationships between latent constructs [156]. For each factor, I use multi-item measurement scales to control for measurement error [85]. The model fit indices reflected excellent

<sup>2</sup>Lavaan package: <https://cran.r-project.org/web/packages/lavaan/index.html>

<sup>3</sup>For alpha, >.70 is acceptable, >.80 is good, >.90 is excellent.

fit<sup>4</sup>:  $\chi^2(384) = 1079.936$ ,  $p < .01$ ; RMSEA = 0.059, 90% CI: [0.055, 0.063], CFI = 0.973, TLI 0.969.

Results of the SEM analysis are reported starting from the left and going towards the right of the model. Only direct relationships and significant results are discussed for the sake of conciseness.

Figure 5.10 depicts the results of the final model.

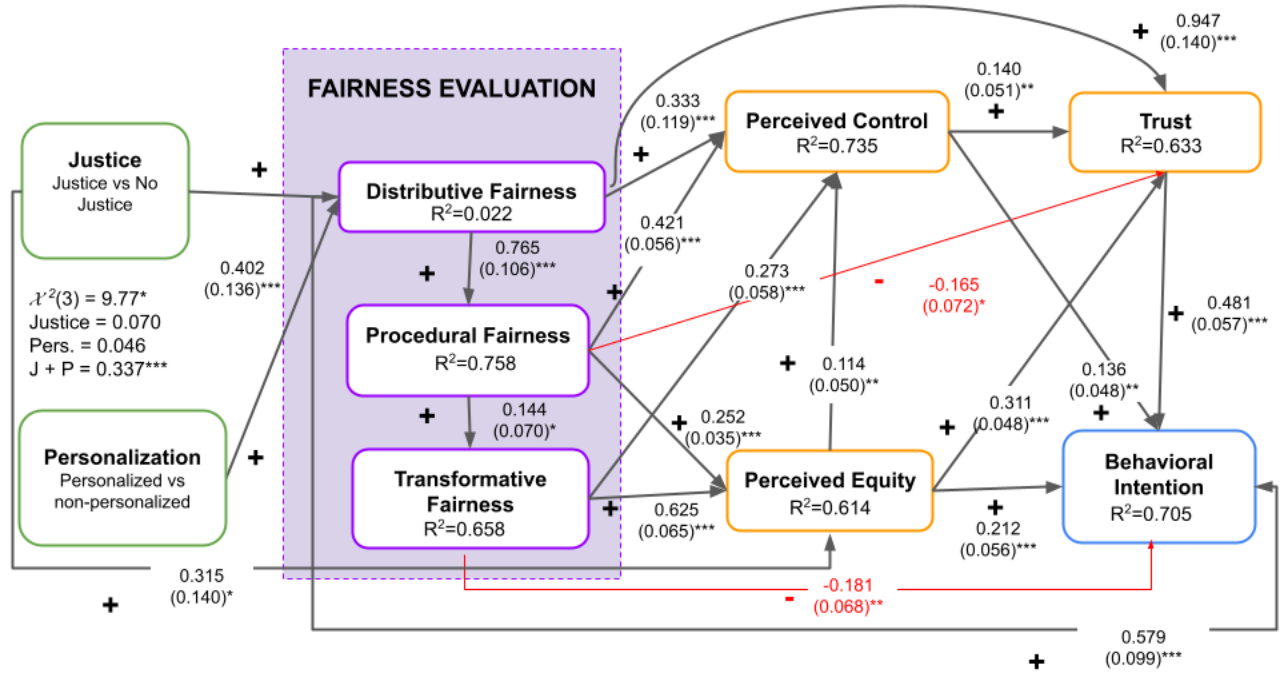


Figure 5.10: The figure above displays the SEM model. Positive relationships are depicted by black arrows versus negative relationships which are depicted by red arrows. The model is color-coded based on the type of latent variable. Green denotes manipulations, purple denotes subjective aspects around fairness, orange denotes subjective aspects about the system, and blue denotes outcomes. Pers. is an abbreviated form of personalization. J+P represents the combined effect of justice and personalization.

The results confirm that justice-oriented countermeasures have a significant effect on perceptions of fairness but only if they are personalized ( $\beta = 0.337$ ,  $p < 0.001$ ) (**H1a and H1c supported**). On average, participants who were exposed to justice-oriented countermeasures (C1 and C3) perceived the system to be performing in a more fair manner compared to the baseline. The results of the Multiple Indicators Multiple Causes (MIMIC) analysis indicate that, all things considered, participants exposed to justice-oriented countermeasures reported higher levels of fairness across all

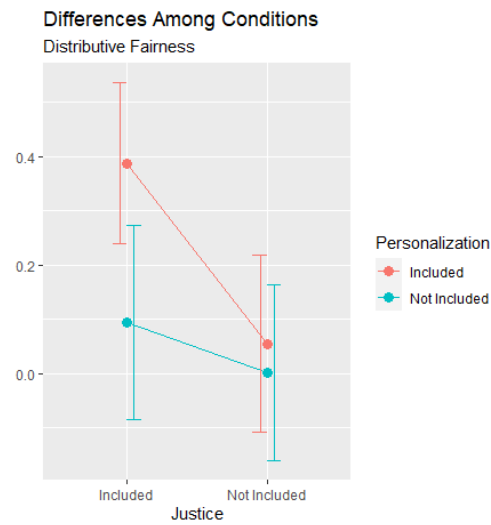
<sup>4</sup>A model should not have a non-significant  $\chi^2$ , but this statistic is regarded as too sensitive [21]. Hu and Bentler [80] propose cutoff values for other fit indices to be: CFI > .96, TLI > .95, and RMSEA < .05, with the upper bound of its 90% CI below 0.10.

three facets. Although there are increased levels of reported fairness if exposed to justice-oriented countermeasures, this effect is only positive with the combined effect of personalization. In other words, participants perceived the system to be more fair if the countermeasures included justice-oriented designs that acknowledged their own personal context. Specifically, participants in C1 rated significantly higher levels of distributive ( $\beta = 0.402, p < 0.01$ ), and transformative fairness ( $\beta = 0.310, p < 0.05$ ). These differences are illustrated in figures 5.11a, 5.11b, and 5.11c.

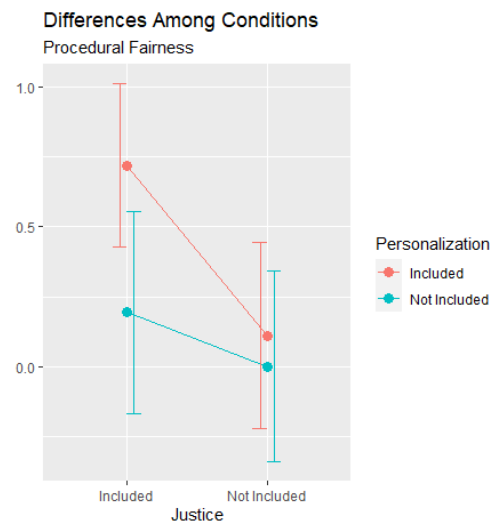
In turn, the effects on fairness perceptions directly influence respondents' attitudes about their interaction with the system. Compared to the baseline, participants with countermeasures that have more detail about the justice outcomes available in their country are more likely to consider the system to be fair ( $\beta = 0.765, p < 0.001$ ) with higher levels of perceived control ( $\beta = 0.333, p < 0.001$ ). Consequently, providing more transparency into the process to seek justice significantly enhances participants' beliefs in their ability to control their justice outcome ( $\beta = 0.421, p < 0.001$ ) while enhancing perceptions of equity ( $\beta = 0.252, p < 0.001$ ). Participants exposed to justice-oriented designs are also more likely to regard the system as more equitable ( $\beta = 0.315, p < 0.05$ ) **(H1b supported)**.

Likewise, having the opportunity to repair a harm via non-punitive means increased participants' belief in their ability to control their justice-related outcomes ( $\beta = 0.273, p < 0.001$ ) **(H2b supported)**. As such, higher levels of perceived control is associated with higher levels of trust in the social media site **(H2a supported)**. However, special attention should be placed on procedural fairness, perceived control, and trust. Countermeasures with lower levels of perceived procedural fairness (not clearly outlining the steps towards justice) does not have an overall total effect on Trust. However, despite the direct negative effect, higher levels of transformative fairness contributes to more positive intentions to adopt countermeasures as there is an overall total effect.

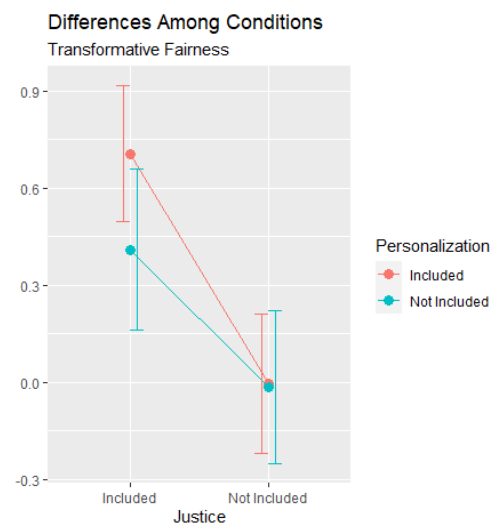
may negatively impact attitudes related to trust ( $\beta = -0.165, p < 0.05$ ). Differences were investigated though a MIMIC analysis and are summarized and illustrated in figures 5.12b, 5.12c, and 5.12a.



(a) Distributive Fairness



(b) Procedural Fairness



(c) Transformative Fairness

Figure 5.11: Marginal effects of the manipulations on perceptions of distributive, transformative, and procedural fairness

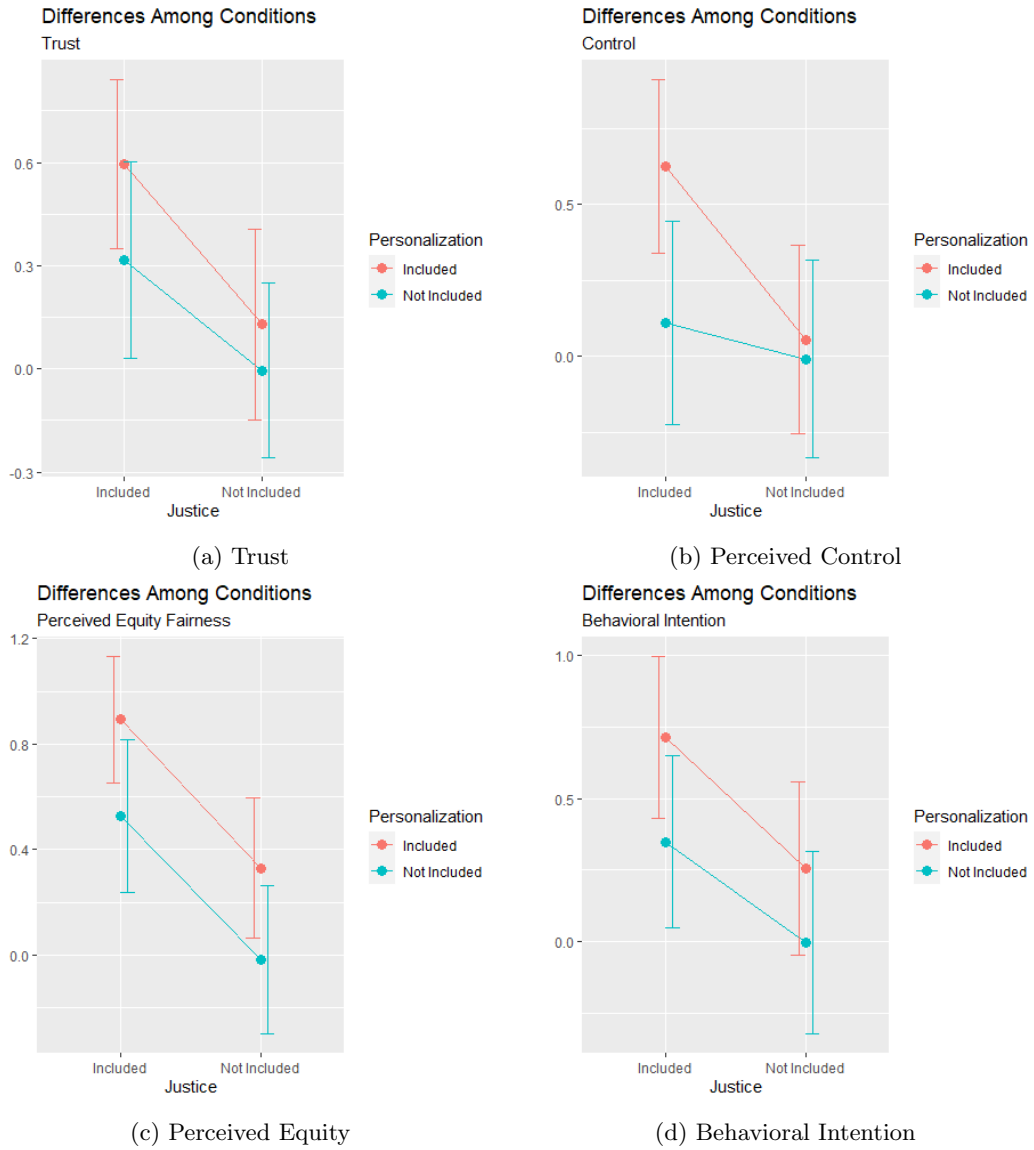


Figure 5.12: Marginal effects on the subjective and outcome factors

### 5.4.3 RQ3: How could justice-oriented countermeasures influence the adoption of protective behaviors within online communities?

Having explored the effects on fairness evaluation and subjective system aspects, it is equally important to investigate the effect on users' intention to adopt the proposed countermeasures. In this study, behavioral intention is operationalized as someone's intention or willingness to adopt a system.

The results indicate that participants' willingness to use the countermeasures has multiple influences. Persons who to have more control over their justice outcome ( $\beta = 0.140$ ,  $p < 0.01$ ) would have increased levels of trust and thus be more willing to adopt the countermeasure ( $\beta = -0.1481$ ,  $p < 0.001$ ). As such, this provides support for **H3a** and **H3b**. Among the conditions, participants with personalized justice-oriented designs are more likely to be willing to adopt the countermeasures ( $F(3) = 10.06$ ,  $p = .018$ ). Differences among all conditions could be observed in figure 5.12d.

Notably, there were also significant country-to-country differences. Across conditions, participants from St. Lucia are more likely to adopt countermeasures ( $F(6) = 22.37$ ,  $p = .001$ ). Figure 5.13 illustrates the country differences related to behavioral intention. In relation to perceived equity, on average, participants from Jamaica were more likely to have higher levels of perceived equity ( $F(3) = 8.38$ ,  $p < .05$ ). Figure 5.14. There were no significant country-to-country differences for any particular condition across all factors.

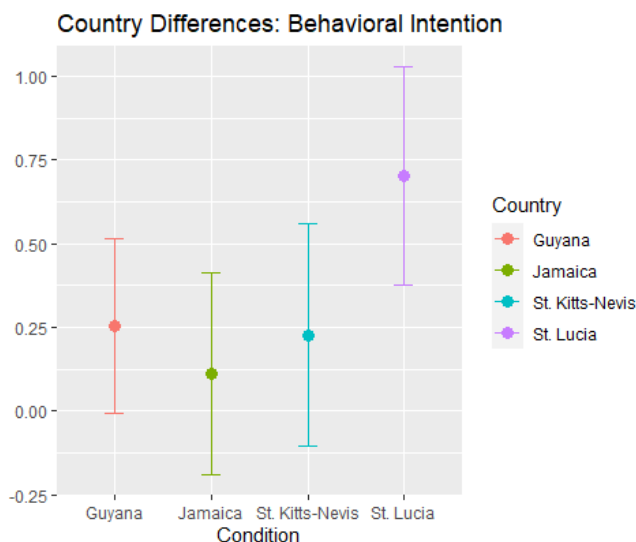


Figure 5.13: Country-to-Country differences among conditions for behavioral intention

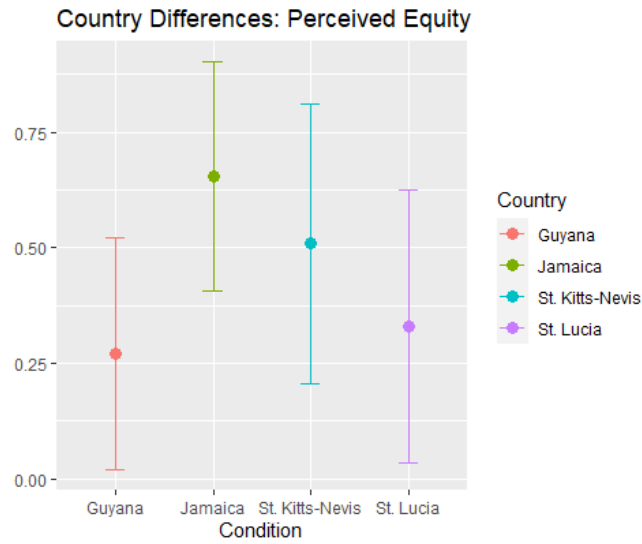


Figure 5.14: Country-to-Country differences among conditions for perceived equity

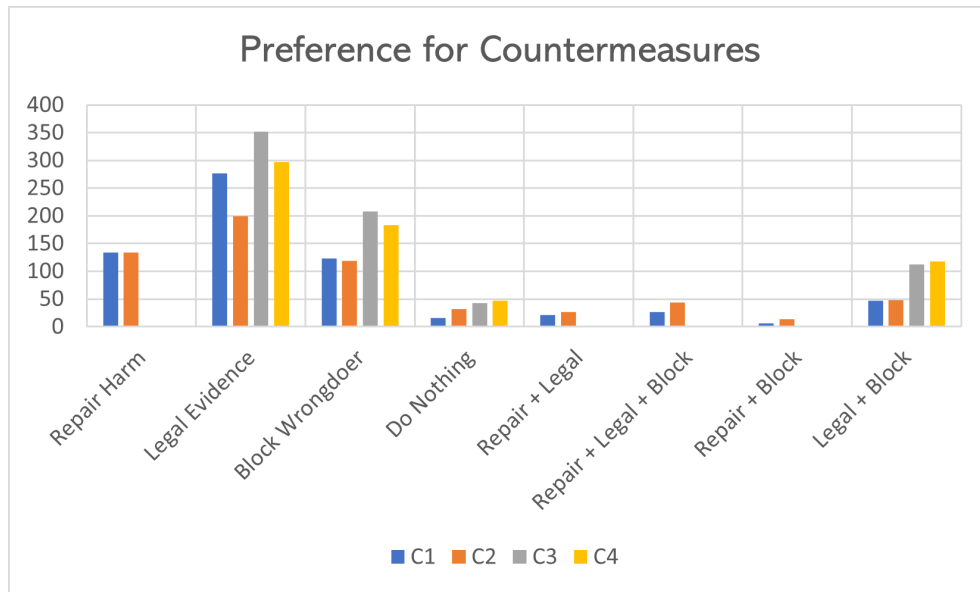


Figure 5.15: Distribution of users' chosen countermeasure/s to respond to harm. Totals are reflective of all four conditions: C1 (Personalized and Justice-Oriented), C2 (Non-personalized and Justice-Oriented), C3 (Not Justice-Oriented but Personalized), C4 (Not Justice-Oriented and Non-personalized).

Beyond general indicators of participants' willingness to adopt, the study also investigated which countermeasures are most preferred. Figure 5.15 demonstrates the distribution of users' chosen countermeasures across all four conditions. Among all conditions, saving an interaction as



legal evidence was the top preference followed by blocking a wrongdoer, then repairing the harm. Among possible combinations of actions, the most preferred options was to save the interaction as potential legal evidence then block then wrongdoer.

Hypothesis	Description	Result
H1a	Justice will directly influence the system’s fairness evaluation.	Supported
H1b	Justice will directly influence perceptions of equity.	Supported
H1c	Personalization will positively influence the system’s fairness evaluation.	Supported
H2a	Fairness evaluation (distributive, procedural, and transformative fairness) will be positively associated with increased levels of trust.	Supported
H2b	Fairness evaluation will be positively associated with increased levels of perceived control.	Supported
H2c	Fairness evaluation will be positively associated with perceived equity.	Partially Supported
H3a	Perceived control will positively be associated with behavioral intention.	Supported
H3b	Trust in SNS is positively associated with behavioral intention.	Supported

Table 5.2: The table above describes the summary of findings related to the hypotheses testing.

#### 5.4.4 Qualitative Responses

Participants responded to free response questions about their experience interacting with the countermeasures. Considering the data was from free responses, the topics would vary versus if the data was collected in a more structured manner such as following a guided interview script. As such, thematic analysis was chosen to explore themes as they emerged from the data [31] and MaxQDA was used to complete the analysis. There were four main themes that emerged: *protection*, *responsiveness*, *awareness*, and *support*.

The most prevalent theme was related to the site providing *protection*. This was most prevalent in Guyana and least prevalent in St. Kitts-Nevis. This theme included discussions on the extent to which people felt the site provided options to effectively address harms. P290 (Guyana, C3) noted *”I consider these solutions to be a first step towards an ultimate goal of protecting user’s information and dissuading other’s from causing harm on such a platform”*. Similarly, participants considered the designs to be *”fairly realistic and ensures that the victims know what they can do to overcome the issue or protect themselves”* (P443, St. Kitts-Nevis, C1). Under this theme, the most

prevalent code was related to legal evidence. Respondents expressed appreciation for the ability to save interactions as potential legal evidence. This was expressed across all conditions. Even if persons did not have contextual information about how to pursue legal justice in their country, the option to save potentially problematic interactions as proof of harm was perceived as useful. One participant mentioned they liked *"...the fact that I can get a recorded documentation showing proof of the criminating evidence that those parties has done something to me"* (P7, St. Lucia, C4). In a similar sense, the countermeasures provided reassurance as participants mentioned that the options helped in understanding the paths available for justice: *"I like that the website had a choice to save post on file which would be legal as a document as you know persons can delete posts. It was also good to know what would happen if a certain path was taken"* (P23, St. Lucia, C2). However, concerns about how legal protections are regarded and enforced locally raised doubts about the effectiveness of saving evidence. *"How safe one may feel after saving for evidence and reporting it depends on how serious the judiciary or law enforcement in their area views the crime"* (157, Jamaica, C3).

Participants also highlighted both affordances and challenges related to the *responsiveness* of the different countermeasures. Although providing insights about potential judicial protections was considered to be valuable, one of the challenges would be the time it takes to pursue that route. As such, participants categorized the appropriateness of the responses. *"Blocking is instant. Legal action is protection on a long-term"* (P152, Guyana, C2). Beyond response time, having flexibility in the ability to respond to harms with varying levels of severity was a valuable attribute. When asked about aspects of the site that they liked, participants expressed the fact *"that they allow both serious repercussions as well as light ones to facilitate different people. Also asking you before affecting the [other] person negatively"* (P225, Guyana, C4). P392 also contributed to this point by saying *"the solutions were appropriate for the scenarios. They offered an opportunity to correct minor offences such as 'spreading misinformation', and to combat major offences such as 'leaking banking information'"* (P392, Jamaica, C2). In a similar sense, the ability to have inaction as a response enhanced attitudes related to control. *"The solutions given were set up to show that you can deal with the problem as you see fit [whether] it be severe or just turn a blind eye"* (P351, Jamaica, C1).

One of the benefits of being exposed to personalized designs was increased *awareness* of local rights. Countermeasures were adjusted to reflect the legal protections in each respective country. As a result, participants acknowledged that an aspect they valued was increased awareness of rights.

P491 expressed that they most appreciated learning *"new information - I did not know of some of the fines and prison time associated with the violations in SKN"* (P491, St. Kitts-Nevis, C3). Interestingly, the presence of judicial information also provoked thoughts around self-regulation and awareness of potentially harmful behaviors to others. *"The option is gear[ed] towards allowing me to act right accordingly to law and my own safety"* (P362, Jamaica, C2).

Lastly, responses related to *support* referred to the extent to which the site provided an environment that felt comfortable and healing for victims. *"Its a refreshing change, and it provides comfortable language to let the victim know their options."* (P136, Guyana, C4). Notably, participants also mentioned valuing the ability to have conflict resolution between themselves as the victims and the wrongdoer as a chance for redemption and being heard: *"[the] approach to resolution two pronged- gives the perpetrator a chance to correct while also the wronged an opportunity to be heard."* (P429, St. Kitts-Nevis, C1).

These insights help to better contextualize the results of the survey. In the next section, I discuss the implications of these results for multiple stakeholders.

## 5.5 Discussion

This study illustrates the effect of justice-oriented countermeasures. Consequently, the results provide theoretical, empirical, and practical contributions. The application of a justice-oriented approach in interaction design presents an avenue for people to have actionable methods to achieve outcomes that they perceive to be fair. Although the empirical findings reveal support for this direction, this final section reflects on underlying design principles and strategies that could serve as guiding goals for designers. The key principles highlighted focus on designing for *awareness*, *accountability*, and *allocation*. The section closes with a discussion of practical implications specifically for Caribbean stakeholders.

### 5.5.1 Designing for Awareness

This study found that incorporating a rights-affirming focus to design enhanced awareness of protections and options for justice. Acknowledging users' varied needs and individual differences has been raised as a major concern in understanding how harms in online spaces could be addressed in a way that is fair but also equitable. Moreover, offering personalized details on how to achieve

justice aims to empower users by arming them with rights-affirming knowledge. This approach is beneficial to those who have experienced harm and those who have not.

Including details on protections provide a sense of safety for victims whereas this information simultaneously contributes to self-regulation. As persons are more aware of potential repercussions they regulate their own behavior to remain within the rules and avoid consequences. Moreover, prior work has argued that the inclusion of more details around how to achieve justice improves users' likelihood of adoption. The results show that there is not an overall effect on procedural fairness. As such, this work builds on understanding the impact of applying justice theories to design while also investigating how those designs inform the transformation of protective tools.

### 5.5.2 Designing for Accountability

Holding wrongdoers responsible involves designing opportunities to immediately protect those affected but also creating avenues for the transformation of perpetrators to reduce recidivism rates. Although prior works have discussed criminal justice theories from a conceptual standpoint, very few studies within HCI spaces have considered applying these principles to inform design artifacts. This work carefully considers how key principles from multiple theories could work in unison to improve accountability. The language included in the designs were carefully chosen and tailored to adopt a respectful survivor focus. The goal of this approach was to highlight the platform acknowledges *the wrongdoing* and *the wrongdoer* all while making the options to hold them responsible very clear. This was received positively as participants expressed their appreciation for language that they deemed as comforting for victims.

Beyond language, the types of actions that should be included as countermeasures have long been argued by different schools of thought. Recent work has argued for the inclusion of restorative justice principles in opposition to punitive approaches but insights were not applied to create or test any designs based on these principles [199, 157]. This research builds on and confirms key assumptions within literature in this area. First, people are open and willing to adopt approaches that are not solely punitive. Current designs of safety countermeasures deployed in the wild are largely dependent on approaches that bans or limits users or content. Although punitive approaches such as blocking are dominant, providing opportunities to repair harm enhances safety perceptions. It should be noted that providing these options does not eliminate the use of punitive approaches but rather serves as an additional option. The results provide support for alternative

methods such as saving problematic interactions as evidence but also the combination of approaches to effectively address issues. There are also temporal considerations as more punitive options may provide immediate relief whereas transformative or judicial options may require more time to achieve a fair outcome. Instead of opposing any particular method, future work could investigate the balance of providing both opportunities for justice in a manner that is not overwhelming.

### **5.5.3 Designing for Allocation**

Allowing persons to have the resources they need would require deep thought about departing from universalist design and moving towards allocating appropriate resources more equitably. From a theoretical standpoint, the proposed conceptual model contributes to HCI and social justice communities by considering how models of criminal justice could inform the design of safety countermeasures. Principles of distributive justice could be applied in technical and non-technical aspects of research. From a technical standpoint, providing more personalized information on how to achieve justice helps to improve overall perceptions of system fairness and helping people feel like they could better control and maintain their safety in online spaces. Thus, designers could reflect on structural inequalities that hinder social justice. This may require incorporating political and cultural considerations to develop effective solutions to robustly address harms [54]. The results demonstrate support for the effectiveness of providing insights about local options to pursue judicial justice. In similar light, designers and developers could examine different resources that may be employed to maintain safety. This may not be limited to judicial and procedural insights but distributing more resources for types of harms that are particularly more severe. For example, it might be more appropriate to allocate more detailed countermeasures if one’s banking information is distributed without consent versus if one interacts with content that would be regarded as disinformation. From a non-technical perspective, increasing diversity in the backgrounds of those involved in the design and development of countermeasures could contribute to a deeper understanding of issues and how to address them effectively.

### **5.5.4 Practical Considerations and Broader Impact**

Thus far, the results have provided support for the consideration of alternative justice-oriented countermeasures in online spaces. However, its deployment would be associated with chal-

lenges that should be carefully considered to avoid unintentionally causing more harm to people who are already victims. As uncovered in chapter 4, even within the Caribbean region there is significant variance in the extent of legislative protections available. Thus, providing reliable and accurate information would be technically challenging and even lead to greater feelings of inequity if more protections are available to others while causing more grief for some. Consequently, designers, researchers and developers could consider aspects of design that could be personalized to provide increased contextual information about local justice options. Also, providing opportunities to connect with local resources to serve as an updated source for this content. For example, across the Caribbean region, GetSafeOnline.org<sup>5</sup> supplies information on how Caribbean citizens could remain safe in online spaces. This would also offer local advocates an opportunity to amplify supportive services which would provide more venues for further impact.

In terms of broader impact, having an understanding of how to design mechanisms for a socially diverse region like the Caribbean could be directly beneficial for geographical areas with similar socioeconomic status and challenges with resources. Based on the foundations explored in this research, designer, developers, and practitioners could consider how to scale safety resources that acknowledge imbalances in resources that would affect fair outcomes.

## 5.6 Conclusion

This chapter began by describing alternative justice-oriented countermeasures and arguing that its inclusion in online space would assist in the people's perceptions of fairness, equity and trust. The empirical findings provide support for this position. Particularly, personalized countermeasures that offer details on localized contextual information on the justice process outperforms all other designs. The results suggest that participants were most willing to adopt countermeasures where they were allowed to save problematic interactions as evidence of harm. Additionally, participants were also willing to adopt countermeasures that allowed them to resolve and repair harm while still being able to apply approaches such as blocking to immediately cease harmful interactions. Overall, this work provides further understanding on how designs could better incorporate fairness, equity and trust in users' pursuit of justice in online spaces.

---

<sup>5</sup>Get Safe Online offers content on how to remain safe in online spaces: <https://www.getsafeonline.org/get-safe-online-around-the-world/>

Table 5.1: The factors of personal characteristics with the Average Variance Extracted (AVE) and the consistency coefficients (Cronbach's  $\alpha$ ), and the items per construct with item factor loadings. Removed items are colored in grey

Factor	Items	Loading
Distributive Fairness AVE: 0.795 Cronbach's $\alpha$ : 0.86	Overall, the solutions proposed by the system were fair.	0.763
	The way the system helped me resolve the problems made me feel like I did not get what I deserved.	
	In resolving the problems, the system gave me what I needed.	0.829
	The proposed solutions I received were not right.	
	The system considered my needs in proposing solutions.	0.771
	The system considered the extent of my effort in resolving the problem.	0.815
Procedural Fairness AVE: 0.813 Cronbach's $\alpha$ : 0.92	I was allowed a great deal of participation in resolving the problem.	0.811
	The solutions proposed by the system were adequately explained to me.	0.717
	I had a great deal of input into the process of resolving the problem.	0.846
	I was able to significantly influence my decision regarding which solution to pursue to resolve the problem.	0.830
	A reasonable rationale was provided for each proposed solution.	0.826
	I had a great deal of control over my decision to resolve the problem.	0.838
Transformative Fairness AVE: 0.753 Cronbach's $\alpha$ : 0.85	The solutions proposed by the system have the potential to repair my relationship with the wrongdoer (the person who caused the problem).	
	The solutions proposed by the system could help wrongdoers cease their harmful behaviors.	0.777
	The system proposed solutions that allowed me to reach an understanding with the wrongdoer.	0.735
	The system offered solutions that had an equal concern toward healing the lives of both those who have been harmed and those who caused harm.	0.782
	The system helps wrongdoers to accept responsibility for their actions.	0.798
	The system gives Community members an active voice in defining justice for victims.	
Perceived Equity AVE: 0.807 Cronbach's $\alpha$ : 0.89	The system proposed solutions that reflect my culture, ethnicity, and identity.	0.766
	The system proposed solutions that are inclusive of individuals within my community.	0.795
	The system offers more options for those who need more help.	0.814
	The system offers resources that would help Caribbean people in particular.	0.837
	The system acknowledged different needs relevant to my culture, ethnicity, and identity.	0.818
Perceived Control AVE: 0.596 Cronbach's $\alpha$ : 0.88	When specifying my preferences for appropriate solutions, I felt I was in control.	0.900
	I think that I had a lot of control over the resolution process.	0.921
	The way I indicated my choice for a resolution made me feel I was in control.	0.883
	I became familiar with the system very quickly.	
	The system helped me to make decisions faster.	
Trust AVE: 0.816 Cronbach's $\alpha$ : 0.88	This social media site is trustworthy.	0.809
	This social media site wants to be known as one who keeps promises and commitments.	0.748
	I trust this social media site keeps my best interests in mind.	0.827
	I find it necessary to be cautious with this social media site.	
	This social media site has more to lose than to gain by not delivering on its promises.	
	This social media site's behavior meets my expectations.	0.874
Behavioral Intention AVE: 0.826 Cronbach's $\alpha$ : 0.90	Assuming I had access to this system, I intend to use it.	0.828
	Given that I had access to this system , I predict that I would use it.	
	I feel quite certain of the benefits I could expect to get if I used this system.	0.845
	To the extent possible, I would use this social media site frequently.	0.856
	Using this new product/service would allow me to do things that I can't easily do now.	0.772

## Chapter 6

# General Conclusions and Future Directions

Algorithmic and data-driven technologies have persistently evolved to become an ever-present force in the daily lives of billions of people. The deployment of these technologies has a tendency to flow from economies of the global north to those of the global south and the lingering threats associated with their use travels across borders as well. However, approaches to achieve fair outcomes in ADDTs vary significantly across borders. This dissertation provides empirical evidence to highlight the threats that plague Caribbean citizens and establish a deep understanding of the current ecosystem that is in place to effectively mitigate these threats. An overview of the dissertation, framed within the socio-ecological model mentioned in Chapter 1, is presented in Figure 6.1.

Through multiple methodological approaches, this dissertation offers four major types of contributions to a range of scholarly communities.

- **Theoretically:** In Chapter 3, I apply the concepts behind Protection Motivation Theory (PMT) to extend knowledge on how Caribbean people perceive, evaluate, and mitigate threats to their online safety.
- **Empirically:** In Chapter 3, I conduct a regional survey on online safety within the Caribbean, which contributes to the limited body of existing HCI research on this population and towards knowledge on the prevalence of threats region-wide.



- **Comparative Analysis:** In Chapter 4, I contribute towards an understanding of the operation of regional law and legal systems and its impact on the formulation of policy related to online safety.
- **Artifact:** In Chapter 5, I contribute a novel approach to the HCI community by applying justice-oriented principles in the design of safety countermeasures.

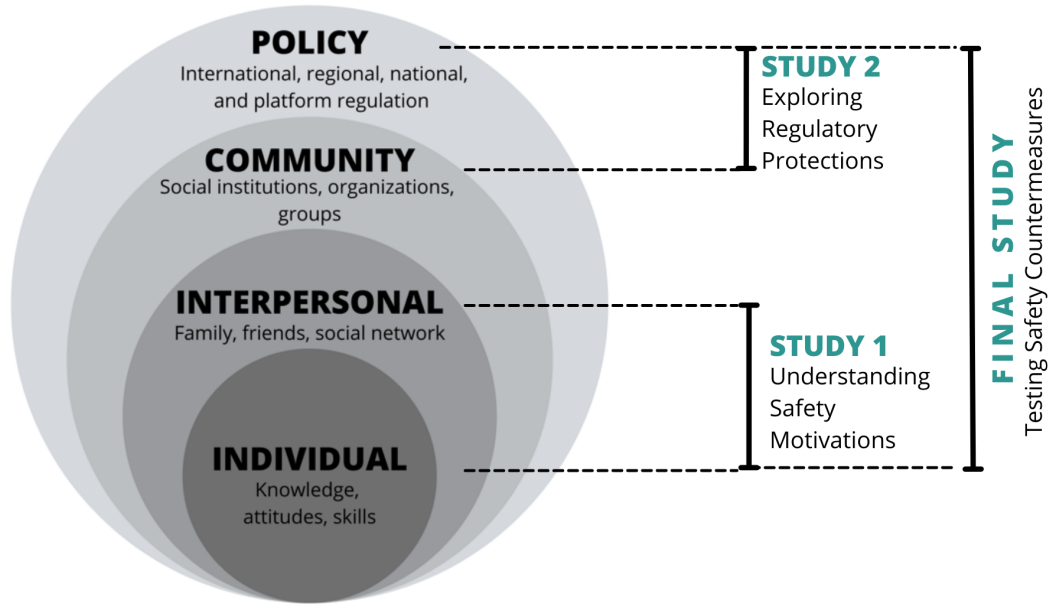


Figure 6.1: Illustration of the different phases of this dissertation

By applying a socio-technical lens, the insights point to a deeper understanding of the concerns regarding online safety by investigating non-technical factors that influence the effectiveness of technical ones. Specifically, the outcomes of the work provide insight into: (1) how people experience threats within ADDTs, (2) the extent to which existing safety tools help them feel protected, (3) the extent to which governments mobilize to enforce regulatory protections, (4) and how justice-oriented countermeasures can be designed to address the shortcomings of technical tools and non-technical paths for justice.

In Chapter 3, I uncover Caribbean citizens feel vulnerable to a wide variety of severe online threats. The wider definition of safety adopted in this study allow us to capture a diverse representation of threats that influence people’s sense of safety. Interestingly, safety concerns are not centered towards any single definition of safety such as privacy or security although concerns are often studied

in silos. This methodology was also helpful in understanding differences in users' interpretation of different threats and their subsequent protective behaviors in response to these threats. The results reveal people are open to utilizing online safety mechanisms depending on the type of threats they address. Once threats travel outside of the boundaries of the technical platform, such as in cases of discrimination or physical stalking, the options available become less appealing. For developers this should be of interest as multiple platforms focus on audience regulation and punitive approaches as the primary method to respond to misbehavior. In these cases, reducing or halting communications with the violator might not be enough as the consequences extend beyond solely communication or beyond harm of the primary target by harming people within that person's social circle. For harassment-related threats, how much people trust a platform plays a significant role in whether they would use the safety mechanisms offered.

Outside of technical means of maintaining safety, Caribbean citizens are provided judicial options for pursuing justice, but the legislative approaches are inconsistent. In Chapter 4, I highlight gaps and challenges in the legislative approach to online safety. Although there are shortcomings in the implementation of protective laws, it raises opportunities for designs that are cognizant of the functional components of that system. Thus, I propose a pathway to the design and evaluation of justice-oriented countermeasures that acknowledge the dynamics of fairness and justice in the creation of equitable solutions. The results of chapter 5 provide support for the development countermeasures that are rights-affirming while balancing people's needs for action.

Overall, this dissertation builds on and extends existing works by providing an understanding of what makes people feel vulnerable in online spaces, exploring approaches that currently address those concerns, and evaluating alternative pathways to offer more equitable opportunities to allow more people to arrive at fair outcomes. With these lessons, I outline and discuss opportunities for future work that embed the values of equitable development in the design of safety countermeasures.

## **6.1 Re-imagining Online Justice Futures**

In this section, I argue for a paradigm shift around our approach to minimizing the costs of ADDT use. Corporate entities have developed entire online worlds and have more recently been looking for new frontiers to conquer while assigning clean up crews to address the social, economic, and regulatory costs that arise along their path towards innovation. Empirical evidence points to

these costs being substantial and rampant, thus, important to understand before arriving at a point where it is harder to investigate, reverse, or mitigate costs considering the pace of advancement with algorithmic technologies. Therefore, if we are to alter our trajectory by centering justice and safety, it becomes important to ask: which voices contribute to the architecture of online systems and by extension their protections? Who conceptualizes notions of safe, equitable, and fair views in safety research and in governance? What role could academic research play in influencing these conversations and having broader impact beyond academic circles? Along this line of reflection, future online safety research could explore the following directions.

### 6.1.1 Expanding our Toolkit

This dissertation has highlighted that technology-facilitated threats extend well beyond digital spaces, thus, the response to these threats requires looking beyond solely digital perspectives to understand how to appropriately respond. Moreover, the completed work also shows that multiple societal influences contribute to the development of solutions that are both effective and equitable. This is particularly important for low-resource communities who may have limited access to resources that would allow them to adequately respond to incidents of harm. Maximizing the benefits of advanced technologies from ADDTs should also include thoughts around how to elevate often-forgotten communities by examining gaps in current infrastructures. Therefore, it is valuable to:

- **Acknowledge the Ecosystem:** Going forward, future studies could deeply investigate the proliferation of harm in online spaces by expanding the purview beyond the confines of technical systems. The completed work demonstrates that the inclusion of socio-technical components enhances our understanding of how harms occur, who they affect, and at what cost. This approach assists in exploring mental models related to how harm manifests in both digital and non-digital realms versus studying specific harms in a silo. Further this approach contributes to a greater understanding of protective actions by considering a wider definition of the costs that seeps into our post-digital worlds. In designing and evaluating countermeasures, stakeholders should consider the scope of use to uncover areas where people are exploited and abused either by a system or via a system. This becomes incredibly critical as algorithmic and data intensive technologies are becoming more entangled with critical areas of life (e.g. transportation, health, and justice systems) where inequalities could be widened. As such, the bar for our approach to

research in this area should be raised. Researchers should be encouraged to challenge existing definitions and notions of safety and justice to allow better options for measuring the true costs of using these systems.

- **Acknowledge Positions:** Beyond alternative avenues for measurement, the narrative and approach to design in safety research should also be expanded. Suggesting grand visions for the implementation of technologically advanced methods of protections are important for innovative progress. However, for communities that often lack a consistent supply of resources to support grand visions that might mean thinking ten years ahead when they are already five years behind. To provide opportunities that may be more equitable, future research could ground design directions by carefully considering how we could support the current infrastructure of our audience versus suggesting a complete overhaul. There is value in meeting people where they are especially when developing for different cultural expectations of safety. Recently, scholars have called for more participatory methodologies to overcome blind spots due to positionality while allowing the people who are impacted to drive the design process and encourage collaboration among stakeholders [30, 205]. Moreover, when connecting with a population that has traditionally not often given a voice in research it is important that the narrative is driven by the people and not the ideals of the researcher. For the HCI community, this might mean exploring paths that are not entirely focused on producing a system or tool but looking at new ways to understand how technology hinders key social processes.

### 6.1.2 Developing at Scale

The development of Algorithmic technologies is advancing at pace where it is challenging to keep up with harms left in the path towards rapid AI innovation. In the rush to release new advances, technologies are deployed first with very little oversight about the guardrails that should be in place to curb potential harm. This problem is further exacerbated when a limited view is adopted about how people should be protected. Throughout the completed work, the concept of equity was at the forefront by supporting the varied needs facing different audiences although they use one system. Implementing equitable countermeasures would be challenging from a development perspective and as such this path would require creative methods to explore developing solutions at scale. There are local advocacy groups with deep knowledge around not only regulatory protections

but resources for support that are intentionally hidden online. The discussions with regional experts in chapter 3 highlighted that there are usually strong systems of support available for especially vulnerable groups (e.g. LGBTQ+ community). For the sake of safety for these vulnerable groups I opt not to delve into the specificity of those resources but rather endorse greater connections between technical systems and local groups who are equip with context-specific resources.

### 6.1.3 Responsible Safety Research

Lastly, executing the research within this dissertation required special considerations to ensure the work was inclusive and conducted in a responsible manner. The population that was chosen exhibited high levels of diversity across multiple socio-cultural and socio-economic metrics such as language, colonial history, culture, currency, economic standing. Although the population has a lot to contribute to research, connecting with potential participants has its challenges. Social computing scholars often rely on crowdsourcing platforms such as Prolific or Amazon Mechanical Turk to recruit participants. Hard to reach populations may not be sufficiently represented or available at all via these venues which would require seeking participants outside common recruitment sources. In pursuing alternatives, it is important to have special considerations in place to ensure that in investigating harm researchers are not creating harm themselves. Exemplar considerations taken throughout the completed studies include:

- **Survey instrument:** Vulnerable groups may feel uncomfortable disclosing gender or sexuality related information due to physical safety risks. Survey options should give respondents the agency to choose if they would like to disclose that information. It is also useful to pilot surveys either with local collaborators or a small group of locals to test for sensitivities to avoid offending an entire population.
- **Recruitment:** Alternatives to crowdsourcing platforms include but are not limited to (a) hiring a local marketing or research firm and (b) utilizing online spaces such as social media to recruit. Both of these options have benefits and costs. However, connecting with local organizations could be valuable in guiding the instrument. While connecting via social media may incur lower financial costs it may be more time consuming to gather larger sample sizes.
- **Remuneration:** Context is important to understand what would be most appealing to potential respondents. For example, a significant majority of Caribbean citizens utilize prepaid

cellular accounts thus making mobile credit more attractive versus a gift card to a US-based retailer.

- **Narrative:** Adopting a deficit-based narrative for communities that may have lower resources could be harmful. Being cognizant of this, researchers should explore directions that support and advance current infrastructures to avoid potentially re-enforcing stereotypes.

The pathway towards safer online environments and experiences remains hopeful as there are opportunities for researchers and stakeholders to work collaboratively to identify and mitigate threats to people's sense of safety. In the pursuit of academic investigations, researchers should also be encouraged to explore multiple avenues for dissemination to inform audiences outside of academic communities. As we observe more advances with algorithmic technologies it would be equally important to highlight options that support responsible and safe methods of deployment while amplifying just and fair pathways that could enforce broader impact for safer futures.

# Appendices

## Appendix A Supplemental Materials for Chapter 3

### A.1 Survey Instrument

#### Platform Usage and Frequency

Please indicate whether you currently use or previously used the following social media sites.

Do you ever use:

(Options: never used it, don't use it anymore, haven't used it in a while, I'm using it now)

- Twitter
- Instagram
- Facebook
- Snapchat
- YouTube
- WhatsApp
- Pinterest
- LinkedIn
- Reddit
- Tik Tok
- WhatsApp FM, GB WhatsApp or any modified version of WhatsApp
- Tumblr

#### Trust in Social Media Platforms

Please indicate your level of agreement with the following:

(Options: 7 pt Likert (Strongly Disagree - Strongly Agree) )

- Social media companies would be trustworthy in handling my information media companies would tell the truth and fulfill promises related to the information provided by me trust that online social media companies would keep my best interests in mind when dealing with my information
- Social media companies are in general predictable and consistent regarding the usage of my information
- Social media companies are always honest with customers when it comes to using the information that I would provide



## **Threat Experience**

Have any of these happened to you?

(Options: Yes or No) Order was randomized.

- Your identity being at risk of theft online
- Being a victim of fraud
- Your login information being at risk
- Your information was stolen to create a fake account
- Your information was used without your knowledge
- Your phone was cloned by someone without permission
- Your information was shared with third parties without your agreement
- Your information was used to send you unwanted commercial offers/ads
- Your views and behaviors being misinterpreted by algorithms
- Your information being used in different contexts from the ones where you disclosed it
- A person spreading malicious rumors about you on social media
- A person taking sexual photos of you without your permission and sharing them on social media
- A person insulting or disrespecting you on social media
- A person creating fake accounts and sending you malicious comments through direct messages on social media
- A person sending you unsolicited explicit content (e.g. naked pictures)
- Someone using your information to stalk you online
- Yourself being discriminated against (e.g. in job selection, receiving price increases, getting no access to a service)
- Your reputation being damaged
- Your relationships with friends or family being damaged
- Your personal safety being at risk
- Someone using your information to stalk you in person

## **Open text:**

- In your opinion, what are the biggest threats to your safety online?

- What do you do to defend against online threats?

### **Perceived Vulnerability**

How likely do you think any of these issues will happen to you?

(Options: 7 point Likert anchored from Extremely Unlikely - Extremely Likely)

*See threats under threat experience.*

### **Perceived Severity**

In your opinion, what are the most severe risks connected with disclosure of personal information on social media sites?

(Options: 7 point Likert anchored from Not at all Severe - Very Severe)

*See threats under threat experience.*

### **Response Efficacy**

Please rate your level of agreement with the following statements.

I feel safer on social media If I have the ability to...

(Options: 7 point Likert anchored from Strongly Disagree - Strongly Agree)

- Use Security controls (such as two factor authentication)
- Complete a Security checkup
- Set up Login alert for my social media accounts
- Use Spam filters
- Create a strong password
- Delete a post
- Hide or restrict content from particular friend/connection
- Unfriend/ Remove Connections
- Block/Remove Followers
- Reject friends/ Delete Requests
- Report harassment on the platform
- Report harassment to the authorities (e.g. the police or build a case with a lawyer)
- Seek legal protection from the platform (e.g. privacy policy)

- Report inappropriate content
- Report potentially fake profile (I.e online impersonation)
- Delete offensive comments
- Hide potentially offensive comments/content
- Seek Support (communal/offline e.g. talking to a friend)
- Ask somebody (e.g., friends, family) what I should do
- Perform safety check online

### **Self Efficacy**

Please rate your level of agreement with the following statements.

If I needed to, I believe I could...

(Options: 7 point Likert anchored from Strongly Disagree - Strongly Agree)

*See protective behaviors under response efficacy.*

### **Behavioral Intention**

Please rate your level of agreement with the following statements.

If I feel unsafe online, I plan to. . .

(Options: 7 point Likert anchored from Strongly Disagree - Strongly Agree)

*See protective behaviors under response efficacy.*

### **Demographics**

*Gender:* What gender do you identify with? (Options: Male, Female, Non-binary, Prefer to self-describe, Prefer not to say)

*Age:* What is your age? (Open text field)

*Education:* What is the highest level of school you have completed or the highest degree you have received? (Options: Less than high school degree, High school graduate (high school diploma or equivalent including GED), Some college but no degree, Associate degree in college (2-year), Bachelor's degree in college (4-year), Master's degree, Doctoral degree, Professional degree (JD, MD), Prefer not to say)

*Race:* Choose one or more races that you consider yourself to be: (Options: White, Black or African American, American Indian or Alaska Native, Native Hawaiian or Pacific Islander, East Indian, Hispanic, Kalinago, Two or more races, Prefer to describe)

Table 1: The survey items for the digital security model with item loading, average variance extracted, and Cronbach’s alpha for each factor. Removed items are colored in grey. Trust was measured once across all models since it measured attitudes towards trustworthiness of platforms independent of harm being faced.

Construct	Label	Item	Loading
Threat Experience AVE: 0.721 $\alpha : 0.83$	Identity Theft	Your identity being at risk of theft online	0.700
	Fraud	Being a victim of fraud	0.695
	Login	Your login information being at risk	0.666
	Fake Account	Your information was stolen to create a fake account	0.742
	Stolen Information	Your information was used without your knowledge	0.795
	Cloned	Your phone was cloned by someone without permission	
Perceived Vulnerability AVE: 0.794 $\alpha : 0.87$	Identity Theft	Your identity being at risk of theft online	0.802
	Fraud	Being a victim of fraud	0.757
	Login	Your login information being at risk	0.836
	Fake Account	Your information was stolen to create a fake account	0.778
	Stolen Information	Your information was used without your knowledge	
	Cloned	Your phone was cloned by someone without permission	
Perceived Severity AVE: 0.898 $\alpha : 0.94$	Identity Theft	Your identity being at risk of theft online	0.896
	Fraud	Being a victim of fraud	0.900
	Login	Your login information being at risk	0.903
	Fake Account	Your information was stolen to create a fake account	0.892
	Stolen Information	Your information was used without your knowledge	
	Cloned	Your phone was cloned by someone without permission	
Response Efficacy AVE: 0.872 $\alpha : 0.92$	2FA	Use Security controls (such as two factor authentication)	0.834
	Security Checkup	Complete a Security checkup	0.912
	Login Alert	Set up Login alert for my social media accounts	0.903
	Strong Password	Your information was stolen to create a fake account	0.836
	Spam Filter	Use Spam filters	
Behavioral Intention AVE: 0.880 $\alpha : 0.94$	2FA	Use Security controls (such as two factor authentication)	0.917
	Security Checkup	Complete a Security checkup	0.916
	Login Alert	Set up Login alert for my social media accounts	0.912
	Spam Filter	Use Spam filters	0.762
	Strong Password	Create a strong password	0.882
Trust AVE: 0.815 $\alpha : 0.91$	Trust1	Social media companies would be trustworthy in handling my information	0.828
	Trust2	Social media companies would tell the truth and fulfill promises related to the information provided by me	0.837
	Trust3	I trust that online companies would keep my best interests in mind when dealing with my information	0.804
	Trust4	Social media companies are in general predictable and consistent regarding the users	0.753

Table 2: The survey items for the harassment model with item loading, average variance extracted, and Cronbach’s alpha for each factor. Removed items are colored in grey.

Construct	Label	Item	Loading
Threat Experience AVE: 0.754 $\alpha : 0.81$	Rumors	A person spreading malicious rumors about you on social media	0.773
	Explicit Photos	A person taking sexual photos of you without your permission and sharing them on social media	
	Insults	A person insulting or disrespecting you on social media	0.861
	Ghost Account	A person creating fake accounts and sending you malicious comments through direct messages on social media	0.694
	Unsolicited	A person sending you unsolicited explicit content (e.g. naked pictures)	0.671
Perceived Vulnerability AVE: 0.817 $\alpha : 0.88$	Rumors	A person spreading malicious rumors about you on social media	0.811
	Explicit Photos	A person taking sexual photos of you without your permission and sharing them on social media	
	Insults	A person insulting or disrespecting you on social media	0.865
	Ghost Account	A person creating fake accounts and sending you malicious comments through direct messages on social media	0.849
	Unsolicited	A person sending you unsolicited explicit content (e.g. naked pictures)	0.735
Perceived Severity AVE: 0.856 $\alpha : 0.93$	Rumors	A person spreading malicious rumors about you on social media	0.882
	Explicit Photos	A person taking sexual photos of you without your permission and sharing them on social media	0.804
	Insults	A person insulting or disrespecting you on social media	0.876
	Ghost Account	A person creating fake accounts and sending you malicious comments through direct messages on social media	0.904
	Unsolicited	A person sending you unsolicited explicit content (e.g. naked pictures)	0.810
Response Efficacy AVE: 0.899 $\alpha : 0.96$	Reporting - on platform	Report harassment on the platform	0.894
	Reporting - to authorities	Report harassment to the authorities (e.g. the police or build a case with a lawyer)	0.889
	Privacy Policy	Seek legal protection from the platform (e.g. privacy policy)	0.903
	Hide Comment	Hide potentially offensive comments/content	0.923
	Report Fake Profile	Report potentially fake profile (I.e online impersonation)	0.913
	Delete Comment	Delete offensive comments	0.869
Behavioral Intention AVE: 0.871 $\alpha : 0.944$	Reporting - on platform	Report harassment on the platform	0.897
	Reporting - to authorities	Report harassment to the authorities (e.g. the police or build a case with a lawyer)	0.851
	Privacy Policy	Seek legal protection from the platform (e.g. privacy policy)	0.825
	Hide Comment	Hide potentially offensive comments/content	0.903

Table 3: The survey items for the access and disclosure model with item loading, average variance extracted, and Cronbach’s alpha for each factor. Removed items are colored in grey.

Construct	Label	Item	Loading
Threat Experience AVE: 0.758 $\alpha : 0.83$	3rd Parties	Your information was shared with third parties without your agreement	0.710
	Ads	Your information was used to send you unwanted commercial offers/ads	0.768
	Algorithms	Your views and behaviors being misinterpreted by algorithms	0.786
	Context	Your information being used in different contexts from the ones where you disclosed it	0.767
Perceived Vulnerability AVE: 0.836 $\alpha : 0.87$	3rd Parties	Your information was shared with third parties without your agreement	0.911
	Ads	Your information was used to send you unwanted commercial offers/ads	0.870
	Algorithms	Your views and behaviors being misinterpreted by algorithms	0.715
	Context	Your information being used in different contexts from the ones where you disclosed it	
Perceived Severity AVE: 0.854 $\alpha : 0.91$	3rd Parties	Your information was shared with third parties without your agreement	0.876
	Ads	Your information was used to send you unwanted commercial offers/ads	0.835
	Algorithms	Your views and behaviors being misinterpreted by algorithms	0.841
	Context	Your information being used in different contexts from the ones where you disclosed it	0.862
Response Efficacy AVE: 0.887 $\alpha : 0.94$	Delete Post	Delete a post	0.785
	Hide Problematic Content	Hide or restrict content from particular friend/connection	0.875
	Unfriend	Unfriend/ Remove Connections	0.920
	Block Friend	Block/Remove Followers	0.921
	Reject Friend Request	Reject friends/ Delete Requests	0.924
Behavioral Intention AVE: 0.896 $\alpha : 0.95$	Delete Post	Delete a post	0.831
	Hide Problematic Content	Hide or restrict content from particular friend/connection	0.859
	Unfriend	Unfriend/ Remove Connections	0.933
	Block Friend	Block/Remove Followers	0.912
	Reject Friends	Reject friends/ Delete Requests	0.941

Table 4: The survey items for the offline model with item loading, average variance extracted, and Cronbach’s alpha for each factor. Removed items are colored in grey.

Construct	Label	Item	Loading
Threat Experience	Discrimination	Yourself being discriminated against (e.g. in job selection, receiving price increases, getting no access to a service)	0.624
	Reputation	Your reputation being damaged	0.831
	Relationships	Your relationships with friends or family being damaged	0.819
	Physical In-Person Stalking	Your personal safety being at risk Someone using your information to stalk you in person	0.822 0.677
Perceived Vulnerability	Discrimination	Yourself being discriminated against (e.g. in job selection, receiving price increases, getting no access to a service)	0.739
	Reputation	Your reputation being damaged	0.886
	Relationships	Your relationships with friends or family being damaged	0.847
	Physical In-Person Stalking	Your personal safety being at risk Someone using your information to stalk you in person	0.828 0.742
Perceived Severity	Discrimination	Yourself being discriminated against (e.g. in job selection, receiving price increases, getting no access to a service)	0.841
	Reputation	Your reputation being damaged	0.935
	Relationships	Your relationships with friends or family being damaged	0.867
	Physical In-Person Stalking	Your personal safety being at risk Someone using your information to stalk you in person	0.909 0.880
Response Efficacy	Support	Seek Support (communal/offline e.g. talking to a friend)	0.893
	Advice	Ask somebody (e.g., friends, family) what I should do	0.827
	Safety Check	Perform safety check online	0.846
Behavioral Intention	Support	Seek Support (communal/offline e.g. talking to a friend)	0.939
	Advice	Ask somebody (e.g., friends, family) what I should do	0.898
	Safety Check	Perform safety check online	0.789

## Appendix B Supplemental Materials for Chapter 5

### B.1 Study Scenarios

### B.2 Prompts

Imagine you are on Community and you realize your personal information such as your name, age, address, and political preference has been shared without your consent for political reasons (corresponds with Figure 2).

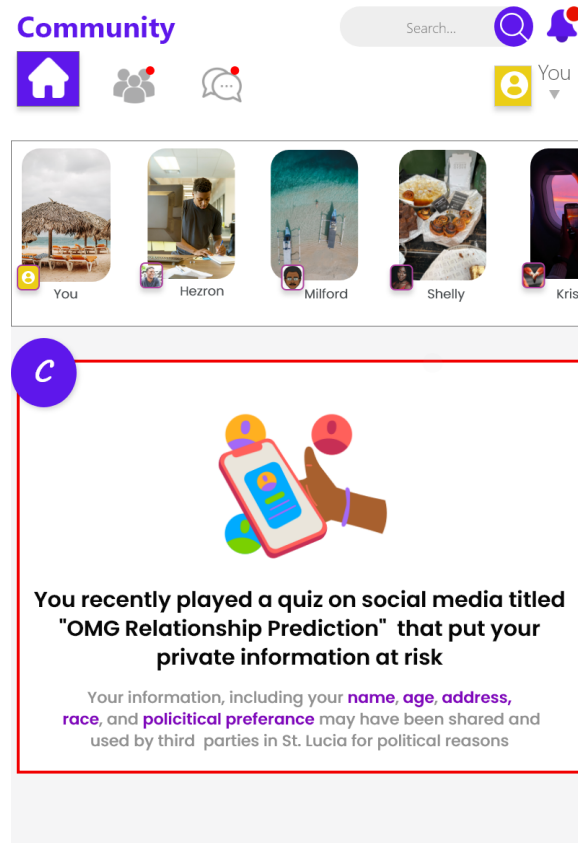


Figure 2: The figure above displays a violation related to access and disclosure of personal information

Imagine you are on Community and you see that you have been hacked and your banking information has been released by someone you know personally (corresponds with Figure 3).



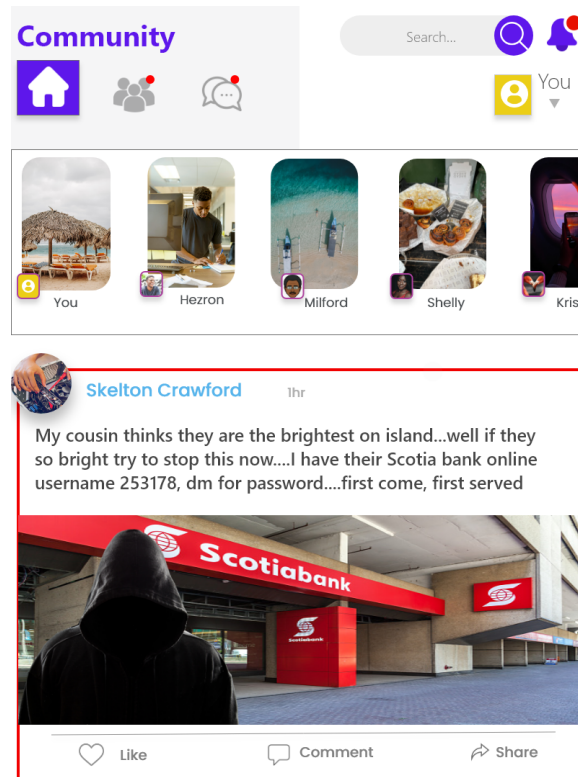


Figure 3: The figure above displays a violation related to unauthorized distribution of banking credentials

Imagine you are on Community and you realize that someone has been stalking your content on Community to follow you in person (corresponds with Figure 4).

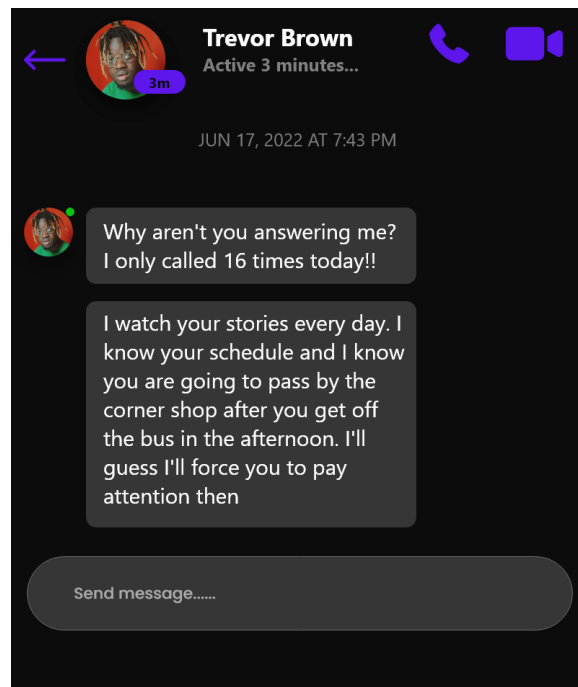


Figure 4: The figure above displays a violation related to online-to-offline threats

Imagine you are on Community and you realize that your old classmate is sharing information that might not be true (corresponds with Figure 5).

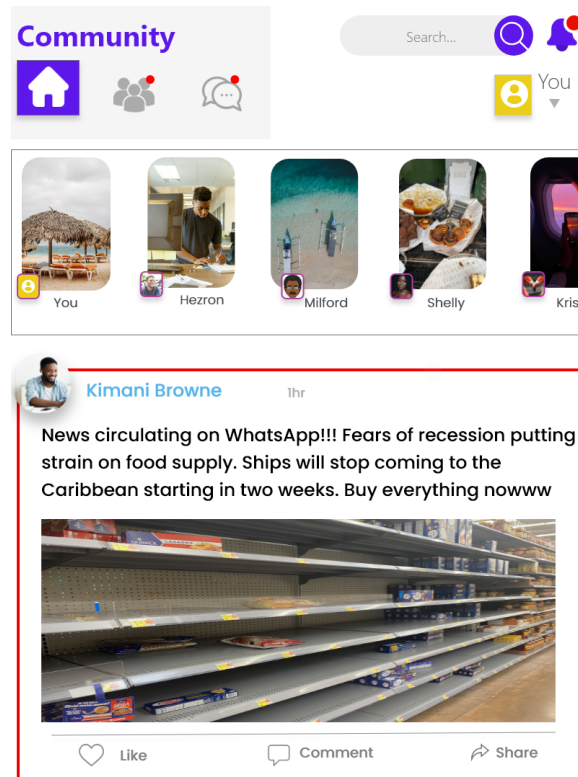


Figure 5: The figure above displays a violation related to misinformation

Imagine you are on Community and you realize that someone who recently fixed your phone is circulating an explicit photo of you that you did not authorize (corresponds with Figure 6).

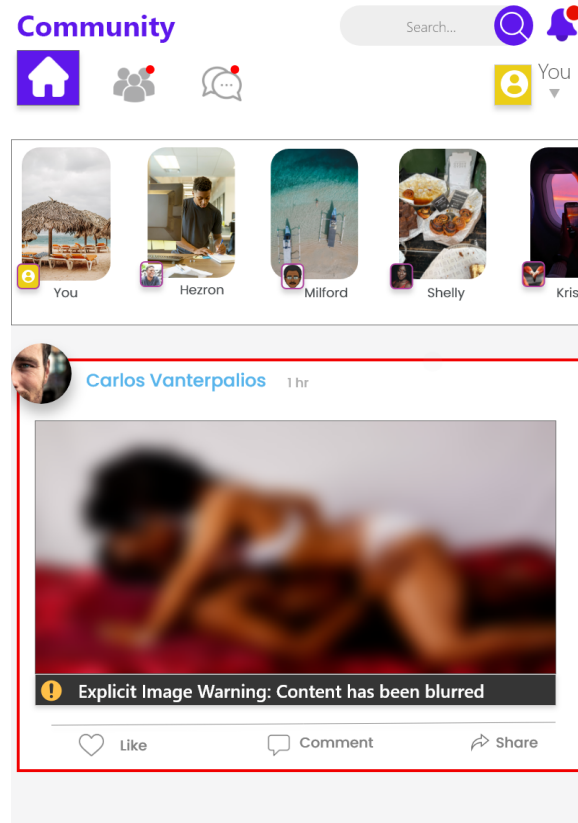


Figure 6: The figure above displays a violation related to the distribution of non-consensual explicit imagery

# Bibliography

- [1] Rediet Abebe, Kehinde Aruleba, Abeba Birhane, Sara Kingsley, George Obaido, Sekou L Remy, and Swathi Sadagopan. Narratives and counternarratives on data sharing in africa. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 329–341, 2021.
- [2] Alessandro Acquisti. Nudging privacy: The behavioral economics of personal information. *IEEE security & privacy*, 7(6):82–85, 2009.
- [3] Alessandro Acquisti and Ralph Gross. Imagined communities: Awareness, information sharing, and privacy on the facebook. In *International workshop on privacy enhancing technologies*, pages 36–58. Springer, 2006.
- [4] JS Adams. Inequity in social exchange. in *advances in experimental social psychology*, vol. 1. berkowitz. new york, ny: Academic press. 1965.
- [5] Yuvraj Agarwal and Malcolm Hall. Protectmyprivacy: detecting and mitigating privacy leaks on ios devices using crowdsourcing. In *Proceeding of the 11th annual international conference on Mobile systems, applications, and services*, pages 97–110. ACM, 2013.
- [6] Ali Abdallah Alalwan, Nripendra P Rana, Yogesh K Dwivedi, and Raed Algharabat. Social media in marketing: A review and analysis of the existing literature. *Telematics and Informatics*, 34(7):1177–1190, 2017.
- [7] David L Alexander, John G Lynch Jr, and Qing Wang. As time goes by: Do cold feet follow warm intentions for really new versus incrementally new products? *Journal of Marketing Research*, 45(3):307–319, 2008.
- [8] Nazanin Andalibi, Oliver L Haimson, Munmun De Choudhury, and Andrea Forte. Understanding social media disclosures of sexual abuse through the lenses of support seeking and anonymity. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 3906–3918, 2016.
- [9] Nazanin Andalibi, Pinar Ozturk, and Andrea Forte. Sensitive self-disclosures, responses, and social support on instagram: the case of# depression. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*, pages 1485–1500, 2017.
- [10] Athanasios Andreou, Giridhari Venkatadri, Oana Goga, Krishna Gummadi, Patrick Loiseau, and Alan Mislove. Investigating ad transparency mechanisms in social media: A case study of facebook’s explanations. 2018.
- [11] Antigua News Room. Husband fined for posting nude photos of wife on social media, January 2019.

- [12] Mariam Asad. Prefigurative design as a method for research justice. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–18, 2019.
- [13] Corlane Barclay. Using frugal innovations to support cybercrime legislations in small developing states: introducing the cyber-legislation development and implementation process model (cyberleg-dpm). *Information Technology for Development*, 20(2):165–195, 2014.
- [14] Corlane Barclay. Cybercrime and legislation: a critical reflection on the cybercrimes act, 2015 of jamaica. *Commonwealth Law Bulletin*, 43(1):77–107, 2017.
- [15] Shaowen Bardzell and Jeffrey Bardzell. Towards a feminist hci methodology: social science, feminism, and hci. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 675–684, 2011.
- [16] Paul Barford, Igor Canadi, Darja Krushevskaja, Qiang Ma, and Shanmugavelayutham Muthukrishnan. Adscape: Harvesting and analyzing online display ads. In *Proceedings of the 23rd international conference on World wide web*, pages 597–608, 2014.
- [17] Solon Barocas and Andrew D Selbst. Big data’s disparate impact. *Calif. L. Rev.*, 104:671, 2016.
- [18] Michael R Bartels. Programmed defamation: Applying sec. 230 of the communications decency act to recommendation systems. *Fordham L. Rev.*, 89:651, 2020.
- [19] Muhammad Ahmad Bashir, Umar Farooq, Maryam Shahid, Muhammad Fareed Zaffar, and Christo Wilson. Quantity vs. quality: Evaluating user interest profiles using ad preference managers. In *NDSS*, 2019.
- [20] Cynthia L Bennett and Os Keyes. What is the point of fairness? disability, ai and the complexity of justice. *ACM SIGACCESS Accessibility and Computing*, (125):1–1, 2020.
- [21] Peter M Bentler and Douglas G Bonett. Significance tests and goodness of fit in the analysis of covariance structures. *Psychological bulletin*, 88(3):588, 1980.
- [22] Morvareed Bidgoli, Bart P. Knijnenburg, and Jens Grossklags. When cybercrimes strike undergraduates. In *2016 APWG Symposium on Electronic Crime Research (eCrime)*, pages 1–10, June 2016. ISSN: 2159-1245.
- [23] Morvareed Bidgoli, Bart P. Knijnenburg, Jens Grossklags, and Brad Wardman. Report Now. Report Effectively. Conceptualizing the Industry Practice for Cybercrime Reporting. In *2019 APWG Symposium on Electronic Crime Research (eCrime)*, pages 1–10, November 2019. ISSN: 2159-1245.
- [24] Reuben Binns, Ulrik Lyngs, Max Van Kleek, Jun Zhao, Timothy Libert, and Nigel Shadbolt. Third Party Tracking in the Mobile Ecosystem. *arXiv preprint arXiv:1804.03603*, 2018.
- [25] Abeba Birhane. Algorithmic injustice: a relational ethics approach. *Patterns*, 2(2):100205, 2021.
- [26] Lindsay Blackwell, Jill Dimond, Sarita Schoenebeck, and Cliff Lampe. Classification and its consequences for online harassment: Design insights from heartmob. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW):1–19, 2017.
- [27] Lindsay Blackwell, Mark Handel, Sarah T Roberts, Amy Bruckman, and Kimberly Voll. Understanding” bad actors” online. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–7, 2018.

- [28] Amelia Bleeker. Creating an enabling environment for e-government and the protection of privacy rights in the caribbean: A review of data protection legislation for alignment with the general data protection regulation. 2020.
- [29] Svetlin Bostandjiev, John O'Donovan, and Tobias Höllerer. Tasteweights: A visual interactive hybrid recommender system. In *Proceedings of the Sixth ACM Conference on Recommender Systems*, RecSys '12, pages 35–42, New York, NY, USA, 2012. ACM.
- [30] Brian Bourke. Positionality: Reflecting on the research process. *The qualitative report*, 19(33):1–9, 2014.
- [31] Virginia Braun and Victoria Clarke. *Thematic analysis*. American Psychological Association, 2012.
- [32] Paula Braveman. Health disparities and health equity: concepts and measurement. *Annu. Rev. Public Health*, 27:167–194, 2006.
- [33] U.S. Embassy Bridgetown. Over ninety media professionals in the eastern caribbean benefit from media and the law training, Sep 2021.
- [34] Moira Burke, Cameron Marlow, and Thomas Lento. Social network activity and social well-being. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1909–1912, 2010.
- [35] Edmond Cambell. Call for revenge porn to be standalone offence. *The Jamaica Gleaner*, 2019.
- [36] CARICOM. Single ict space, cyber security for discussion at ict officials' meeting, Nov 2020.
- [37] James Chalmers and Fiona Leverick. A comparative analysis of hate crime legislation: A report to the hate crime legislation review. 2017.
- [38] Hongliang Chen, Christopher E Beaudoin, and Traci Hong. Securing online privacy: An empirical test on internet scam victimization, online privacy concerns, and privacy protection behaviors. *Computers in Human Behavior*, 70:291–302, 2017.
- [39] Hichang Cho, Jae-Shin Lee, and Siyoung Chung. Optimistic bias about online privacy risks: Testing the moderating effects of perceived controllability and prior experience. *Computers in Human Behavior*, 26(5):987–995, 2010.
- [40] Jonathan Clough. The council of europe convention on cybercrime: defining crime in a digital world. In *Criminal Law Forum*, volume 23, pages 363–391. Springer, 2012.
- [41] Jason A Colquitt. On the dimensionality of organizational justice: a construct validation of a measure. *Journal of applied psychology*, 86(3):386, 2001.
- [42] Sasha Costanza-Chock. Design justice: Towards an intersectional feminist framework for design theory and practice. *Proceedings of the Design Research Society*, 2018.
- [43] Sasha Costanza-Chock. *Design justice: Community-led practices to build the worlds we need*. The MIT Press, 2020.
- [44] Cassie Cox. Protecting victims of cyberstalking, cyberharassment, and online impersonation through prosecutions and effective laws. *Jurimetrics*, pages 277–302, 2014.
- [45] Anne K Cybenko and George Cybenko. Ai and fake news. *IEEE Intelligent Systems*, 33(5):1–5, 2018.

- [46] José de Arimatéia da Cruz and Taylor Alvarez. Small islands, big problems: Cybersecurity in the caribbean realm. *Journal Article— Dec*, 2(7):44am, 2015.
- [47] Erin Daly. Transformative justice: Charting a path to reconciliation. *Int’l Legal Persp.*, 12:73, 2001.
- [48] Amit Datta, Michael Carl Tschantz, and Anupam Datta. Automated experiments on ad privacy settings. *Proceedings on privacy enhancing technologies*, 2015(1):92–112, 2015.
- [49] Fred D Davis and Viswanath Venkatesh. A critical assessment of potential measurement biases in the technology acceptance model: three experiments. *International journal of human-computer studies*, 45(1):19–45, 1996.
- [50] Pieter Walter de Vries. *Trust in systems: effects of direct and indirect information*. Technische Universiteit Eindhoven, 2004.
- [51] Lynn Dombrowski, Ellie Harmon, and Sarah Fox. Social justice-oriented interaction design: Outlining key design strategies and commitments. In *Proceedings of the 2016 ACM Conference on Designing Interactive Systems*, pages 656–671, 2016.
- [52] Vicente Dominguez, Pablo Messina, Ivania Donoso-Guzmán, and Denis Parra. The effect of explanations and algorithmic accuracy on visual recommender systems of artistic images. In *Proceedings of the 24th International Conference on Intelligent User Interfaces, IUI ’19*, pages 408–416, New York, NY, USA, 2019. ACM.
- [53] Charlette Donalds, Corlane Barclay, and Kweku-Muata Osei-Bryson. Cybercrime and cybersecurity in the global south: Concepts, strategies and frameworks for greater resilience, 2022.
- [54] Paul Dourish. Hci and environmental sustainability: the politics of design and the design of politics. In *Proceedings of the 8th ACM conference on designing interactive systems*, pages 1–10, 2010.
- [55] AB Ericsson. Ai: Enhancing customer experience in a complex 5g world. *Ericsson Mobility Report*,. Retrieved 30th July, 2021.
- [56] Motahhare Eslami, Sneha R. Krishna Kumaran, Christian Sandvig, and Karrie Karahalios. Communicating algorithmic process in online behavioral advertising. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI ’18*, New York, NY, USA, 2018. Association for Computing Machinery.
- [57] Alexander Felfernig and Bartosz Gula. An empirical study on consumer behavior in the interaction with knowledge-based recommender applications. In *The 8th IEEE International Conference on E-Commerce Technology and The 3rd IEEE International Conference on Enterprise Computing, E-Commerce, and E-Services (CEC/EEE’06)*, pages 37–37. ieee, 2006.
- [58] Simone Fischer-Hübner, Julio Angulo, Farzaneh Karegar, and Tobias Pulls. Transparency, Privacy and Trust—Technology for Tracking and Controlling My Data Disclosures: Does This Work? In *IFIP International Conference on Trust Management*, pages 3–14. Springer, 2016.
- [59] Luciano Floridi and Josh Cowls. A unified framework of five principles for ai in society. In *Ethics, Governance, and Policies in Artificial Intelligence*, pages 5–17. Springer, 2021.
- [60] Donna L Floyd, Steven Prentice-Dunn, and Ronald W Rogers. A meta-analysis of research on protection motivation theory. *Journal of applied social psychology*, 30(2):407–429, 2000.



- [61] Global Fund for Women. Online violence: Just because it’s virtual doesn’t make it any less real.
- [62] Gerhard Friedrich and Markus Zanker. A taxonomy for generating explanations in recommender systems. *AI Magazine*, 32:90–98, 2011.
- [63] Tarleton Gillespie. *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press, 2018.
- [64] Tarleton Gillespie. Content moderation, ai, and the question of scale. *Big Data & Society*, 7(2):2053951720943234, 2020.
- [65] Stephen W Gilliland. The perceived fairness of selection systems: An organizational justice perspective. *Academy of management review*, 18(4):694–734, 1993.
- [66] Alyssa Glass, Deborah L McGuinness, and Michael Wolverton. Toward establishing trust in adaptive agents. In *Proceedings of the 13th international conference on Intelligent user interfaces*, pages 227–236. ACM, 2008.
- [67] Jamaica Gleaner. Man charged for ‘revenge porn’, accused of posting ex-girlfriend’s nude pics on social media, September 2017.
- [68] Yany Grégoire and Robert J Fisher. Customer betrayal and retaliation: when your best customers become your worst enemies. *Journal of the Academy of Marketing Science*, 36(2):247–261, 2008.
- [69] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):93, 2019.
- [70] David Gunning, Mark Stefik, Jaesik Choi, Timothy Miller, Simone Stumpf, and Guang-Zhong Yang. Xai—explainable artificial intelligence. *Science Robotics*, 4(37), 2019.
- [71] Zach Harned and Hanna Wallach. Stretching human laws to apply to machines: The dangers of a” colorblind” computer. *Fla. St. UL Rev.*, 47:617, 2019.
- [72] A Hasinoff, AD Gibson, and N Salehi. The promise of restorative justice in addressing online harm. *Tech Stream*, 27, 2020.
- [73] Rebecca M Hayes and Molly Dragiewicz. Unsolicited dick pics: Erotica, exhibitionism or entitlement? In *Women’s Studies International Forum*, volume 71, pages 114–120. Elsevier, 2018.
- [74] NM Haynes, CL Emmons, M Ben-Avie, and JP Comer. The school development program student, staff, and parent school climate surveys. *New Haven, CT: Yale Child Study Center*, 2001.
- [75] Raul Herbster, Scott DellaTorre, Peter Druschel, and Bobby Bhattacharjee. Privacy Capsules: Preventing Information Leaks by Mobile Apps. In *Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services*, MobiSys ’16, pages 399–411, New York, NY, USA, 2016. ACM.
- [76] Jonathan L. Herlocker, Joseph A. Konstan, and John Riedl. Explaining collaborative filtering recommendations. In *Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work*, CSCW ’00, pages 241–250, New York, NY, USA, 2000. ACM.

- [77] Steven Hick, Edward Halpin, and Eric Hoskins. *Human rights and the Internet*. Springer, 2016.
- [78] Steve Hoeffler. Measuring preferences for really new products. *Journal of marketing research*, 40(4):406–420, 2003.
- [79] Ming Hsu, Cedric Anen, and Steven R Quartz. The right and the good: distributive justice and neural encoding of equity and efficiency. *science*, 320(5879):1092–1095, 2008.
- [80] Li-tze Hu and Peter M Bentler. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural equation modeling: a multidisciplinary journal*, 6(1):1–55, 1999.
- [81] JANE IM, SARITA SCHOENEBECK, MARILYN IRIARTE, GABRIEL GRILL, DARICIA WILKINSON, AMNA BATOOL, RAHAF ALHARBI, AUDREY FUNWIE, TERGEL GANKHUU, ERIC GILBERT, et al. Women’s perspectives on harm and justice after online harassment. In *Proceedings of the 2022 ACM Conference On Computer-Supported Cooperative Work And Social Computing*, 2022.
- [82] Amnesty International. Toxic twitter - a toxic place for women.
- [83] Costas Iordanou, Nicolas Kourtellis, Juan Miguel Carrascosa, Claudio Soriente, Ruben Cuevas, and Nikolaos Laoutaris. Beyond content analysis: Detecting targeted ads via distributed counting. In *Proceedings of the 15th International Conference on Emerging Networking Experiments And Technologies*, pages 110–122, 2019.
- [84] Lilly Irani, Janet Vertesi, Paul Dourish, Kavita Philip, and Rebecca E Grinter. Postcolonial computing: a lens on design and development. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1311–1320, 2010.
- [85] James Jaccard and Choi K Wan. Measurement error in the analysis of interaction effects between continuous predictors using multiple regression: Multiple indicator and structural equation approaches. *Psychological bulletin*, 117(2):348, 1995.
- [86] Jialun Aaron Jiang, Morgan Klaus Scheuerman, Casey Fiesler, and Jed R Brubaker. Understanding international perceptions of the severity of harmful content online. *PloS one*, 16(8):e0256762, 2021.
- [87] Jialun’Aaron’ Jiang, Skyler Middler, Jed R Brubaker, and Casey Fiesler. Characterizing community guidelines on social media platforms. In *Conference Companion Publication of the 2020 on Computer Supported Cooperative Work and Social Computing*, pages 287–291, 2020.
- [88] Anna Jobin, Marcello Ienca, and Effy Vayena. The global landscape of ai ethics guidelines. *Nature Machine Intelligence*, 1(9):389–399, 2019.
- [89] Allen C Johnston and Merrill Warkentin. Fear appeals and information security behaviors: An empirical study. *MIS quarterly*, pages 549–566, 2010.
- [90] Andrew Jolivet. Research justice: radical love as a strategy for social transformation. *Research justice: methodologies for social change*, pages 5–12, 2015.
- [91] Mudasir Kamal and William J Newman. Revenge pornography: Mental health implications and related legislation. *Journal of the American Academy of Psychiatry and the Law Online*, 44(3):359–367, 2016.

- [92] Antti Kangasrääsiö, Dorota Glowacka, and Samuel Kaski. Improving controllability and predictability of interactive recommendation interfaces for exploratory search. In *Proceedings of the 20th International Conference on Intelligent User Interfaces*, IUI '15, pages 247–251, New York, NY, USA, 2015. ACM.
- [93] Gopinaath Kannabiran, Jeffrey Bardzell, and Shaowen Bardzell. How hci talks about sexuality: discursive strategies, blind spots, and opportunities for future research. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 695–704, 2011.
- [94] Dan J Kim, Charles Steinfield, and Ying-Ju Lai. Revisiting the role of web assurance seals in business-to-consumer electronic commerce. *Decision Support Systems*, 44(4):1000–1015, 2008.
- [95] Pauline T Kim. Data-driven discrimination at work. *Wm. & Mary L. Rev.*, 58:857, 2016.
- [96] René F Kizilcec. How much information? effects of transparency on trust in an algorithmic interface. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 2390–2395, 2016.
- [97] Bart P Knijnenburg, Niels JM Reijmer, and Martijn C Willemsen. Each to his own: how different users call for different interaction methods in recommender systems. In *Proceedings of the fifth ACM conference on Recommender systems*, pages 141–148. ACM, 2011.
- [98] Johannes Knoll. Advertising in social media: a review of empirical evidence. *International Journal of Advertising*, 35(2):266–300, 2016.
- [99] James Konow. Which is the fairest one of all? a positive analysis of justice theories. *Journal of economic literature*, 41(4):1188–1239, 2003.
- [100] Yubo Kou and Xinning Gui. Mediating community-ai interaction through situated explanation: The case of ai-led moderation. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–27, 2020.
- [101] Pigi Kouki, James Schaffer, Jay Pujara, John O'Donovan, and Lise Getoor. Personalized explanations for hybrid recommender systems. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, IUI '19, pages 379–390, New York, NY, USA, 2019. ACM.
- [102] Hanna Krasnova, Sarah Spiekermann, Ksenia Koroleva, and Thomas Hildebrand. Online social networks: Why we disclose. *Journal of information technology*, 25(2):109–125, 2010.
- [103] Hanna Krasnova and Natasha F Veltri. Privacy calculus on social networking sites: Explorative evidence from germany and usa. In *2010 43rd Hawaii international conference on system sciences*, pages 1–10. IEEE, 2010.
- [104] Robert LaRose, Ying Ju Lai, Ryan Lange, Bradley Love, and Yuehua Wu. Sharing or piracy? an exploration of downloading behavior. *Journal of Computer-Mediated Communication*, 11(1):1–21, 2005.
- [105] Mathias Lecuyer, Riley Spahn, Yannis Spiliopoulos, Augustin Chaintreau, Roxana Geambasu, and Daniel Hsu. Sunlight: Fine-grained targeting detection at scale with statistical confidence. In *Proceedings of the 22nd ACM SIGSAC*, pages 554–566, 2015.
- [106] Young Eun Lee and Izak Benbasat. Research note—the influence of trade-off difficulty caused by preference elicitation methods on user acceptance of recommendation agents across loss and gain conditions. *Information Systems Research*, 22(4):867–884, 2011.

- [107] Amanda Lenhart, Michele Ybarra, Kathryn Zickuhr, and Myeshia Price-Feeney. *Online harassment, digital abuse, and cyberstalking in America*. Data and Society Research Institute, 2016.
- [108] Jiawei Li, Qing Xu, Neal Shah, and Tim K Mackey. A machine learning approach for the detection and characterization of illicit drug dealers on instagram: model evaluation study. *Journal of medical Internet research*, 21(6):e13803, 2019.
- [109] Yao Li, Alfred Kobsa, Bart P. Knijnenburg, Carolyn Nguyen, and others. Cross-Cultural Privacy Prediction. *Proceedings on Privacy Enhancing Technologies*, 2017(2):113–132, 2017.
- [110] Yao Li, Alfred Kobsa, Bart P Knijnenburg, M-H Carolyn Nguyen, et al. Cross-cultural privacy prediction. *Proc. Priv. Enhancing Technol.*, 2017(2):113–132, 2017.
- [111] Ting-Peng Liang, Hung-Jen Lai, and Yi-Cheng Ku. Personalized content recommendation and user satisfaction: Theoretical synthesis and empirical findings. *Journal of Management Information Systems*, 23(3):45–70, 2006.
- [112] Brian Y. Lim, Anind K. Dey, and Daniel Avrahami. Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, pages 2119–2128, New York, NY, USA, 2009. ACM.
- [113] Han Lin, William Tov, and Lin Qiu. Emotional disclosure on social networking sites: The role of network structure and psychological needs. *Computers in Human Behavior*, 41:342–350, 2014.
- [114] Jialiu Lin, Shahriyar Amini, Jason I. Hong, Norman Sadeh, Janne Lindqvist, and Joy Zhang. Expectation and Purpose: Understanding Users’ Mental Models of Mobile App Privacy Through Crowdsourcing. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, UbiComp '12, pages 501–510, New York, NY, USA, 2012. ACM.
- [115] Sebastian Linxen, Christian Sturm, Florian Brühlmann, Vincent Cassau, Klaus Opwis, and Katharina Reinecke. How weird is chi? In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2021.
- [116] Emma J Llansó. No amount of “ai” in content moderation will solve filtering’s prior-restraint problem. *Big Data & Society*, 7(1):2053951720920686, 2020.
- [117] Wainer Lusoli, Margherita Bacigalupo, Francisco Lupiáñez-Villanueva, Norberto Nuno Gomes de Andrade, Shara Monteleone, and Ioannis Maghiros. Pan-european survey of practices, attitudes and policy preferences as regards personal identity data management. *JRC Scientific and Policy Reports, EUR*, 25295, 2012.
- [118] May O Lwin, Benjamin Li, and Rebecca P Ang. Stop bugging me: An examination of adolescents’ protection behavior against online harassment. *Journal of adolescence*, 35(1):31–41, 2012.
- [119] Mark Lyndersay. Considering caribbean data protection progress, Aug 2021.
- [120] Michael A Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. Co-designing checklists to understand organizational challenges and opportunities around fairness in ai. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2020.

- [121] Mary Madden, Michele Gilman, Karen Levy, and Alice Marwick. Privacy, poverty, and big data: A matrix of vulnerabilities for poor americans. *Washington University Law Review*, 95:53, 2017.
- [122] James E Maddux and Ronald W Rogers. Protection motivation and self-efficacy: A revised theory of fear appeals and attitude change. *Journal of experimental social psychology*, 19(5):469–479, 1983.
- [123] D Harrison McKnight, Vivek Choudhury, and Charles Kacmar. Developing and validating trust measures for e-commerce: An integrative typology. *Information systems research*, 13(3):334–359, 2002.
- [124] Sean M McNee, Shyong K Lam, Joseph A Konstan, and John Riedl. Interfaces for eliciting new user preferences in recommender systems. In *Proceedings of the 9th International Conference on User Modeling*, pages 178–187, Berlin, Heidelberg, 2003. Springer.
- [125] Maryam Mehrnezhad and Teresa Almeida. Caring for intimate data in fertility technologies. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–11, 2021.
- [126] Vishwali Mhasawade, Yuan Zhao, and Rumi Chunara. Machine learning and algorithmic fairness in public and population health. *Nature Machine Intelligence*, 3(8):659–666, 2021.
- [127] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 2018.
- [128] Norshidah Mohamed and Ili Hawa Ahmad. Information privacy concerns, antecedents and privacy measure use in social networking sites: Evidence from malaysia. *Computers in Human Behavior*, 28(6):2366–2375, 2012.
- [129] Garrett Morrow, Briony Swire-Thompson, Jessica Montgomery Polny, Matthew Kopeck, and John P Wihbey. The emerging science of content labeling: Contextualizing social media content moderation. *Journal of the Association for Information Science and Technology*, 2021.
- [130] Sarah Myers West. Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society*, 20(11):4366–4383, 2018.
- [131] Thomas B Nachbar. Algorithmic fairness, algorithmic discrimination. *Fla. St. UL Rev.*, 48:509, 2020.
- [132] Loop News. Guyanese cop charged for sharing ex-girlfriend’s nudes online | Loop Trinidad & Tobago. *Loop News*, 2018.
- [133] Fayika Farhat Nova, Michael Ann DeVito, Pratyasha Saha, Kazi Shohanur Rashid, Shashwata Roy Turzo, Sadia Afrin, and Shion Guha. ” facebook promotes more harassment” social media ecosystem, skill and marginalized hijra identity in bangladesh. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–35, 2021.
- [134] John O’Donovan, Barry Smyth, Brynjar Gretarsson, Svetlin Bostandjiev, and Tobias Höllerer. Peerchooser: Visual interactive recommendation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI ’08, pages 1085–1088, New York, NY, USA, 2008. ACM.
- [135] Cybercrime Programme Office of the Council of Europe. Report on the regional conference on cybercrime strategies and policies and features of the budapest convention for the caribbean community, Jun 2019.

- [136] UN Office of the High Commissioner. Un experts urge states and companies to address online gender-based abuse but warn against censorship, Mar 2017.
- [137] Sofia C Olhede and Patrick J Wolfe. The growing ubiquity of algorithms in society: implications, impacts and innovations. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376(2128):20170364, 2018.
- [138] Cathy O’neil. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books, 2016.
- [139] Jessica A Pater, Moon K Kim, Elizabeth D Mynatt, and Casey Fiesler. Characterizations of online harassment: Comparing policies across social media platforms. In *Proceedings of the 19th international conference on supporting group work*, pages 369–374, 2016.
- [140] Kimberly Pavlik et al. Cybercrime, hacking, and legislation. *Journal of Cybersecurity Research (JCR)*, 2(1):13–16, 2017.
- [141] Sandra Petronio. Communication privacy management theory: What do we know about family privacy regulation? *Journal of family theory & review*, 2(3):175–196, 2010.
- [142] Sandra Petronio. *Boundaries of Privacy: Dialectics of Disclosure*. SUNY Press, February 2012. Google-Books-ID: 8v89W\_oJQ0wC.
- [143] Leah Lakshmi Piepzna-Samarasinha and Ejeris Dixon. *Beyond survival: Strategies and stories from the transformative justice movement*. AK Press, 2020.
- [144] Pearl Pu and Li Chen. Trust Building with Explanation Interfaces. In *Proceedings of the 11th International Conference on Intelligent User Interfaces*, IUI ’06, pages 93–100, New York, NY, USA, 2006. ACM. event-place: Sydney, Australia.
- [145] Elaheh Raisi. *Weakly Supervised Machine Learning for Cyberbullying Detection*. PhD thesis, Virginia Tech, 2019.
- [146] Ashwin Ram. *Question-driven understanding: An integrated theory of story understanding, memory and learning*. PhD thesis, Yale University, 1989.
- [147] Abbas Razaghpanah, Rishab Nithyanand, Narseo Vallina-Rodriguez, Srikanth Sundaresan, Mark Allman, Christian Kreibich, and Phillipa Gill. Apps, Trackers, Privacy, and Regulators: A Global Study of the Mobile Tracking Ecosystem. 2018.
- [148] Abbas Razaghpanah, Narseo Vallina-Rodriguez, Srikanth Sundaresan, Christian Kreibich, Phillipa Gill, Mark Allman, and Vern Paxson. Haystack: In situ mobile traffic analysis in user space. *ArXiv e-prints*, 2015.
- [149] Angel Marcelo Rea-Guamán, ID Sanchez-Garcia, T San Feliu, and JA Calvo-Manzano. Maturity models in cybersecurity: A systematic review. In *2017 12th Iberian conference on information systems and technologies (CISTI)*, pages 1–6. IEEE, 2017.
- [150] Elissa M Redmiles, Jessica Bodford, and Lindsay Blackwell. “i just want to feel safe”: A diary study of safety perceptions on social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, pages 405–416, 2019.
- [151] Jingjing Ren, Ashwin Rao, Martina Lindorfer, Arnaud Legout, and David Choffnes. ReCon: Revealing and Controlling PII Leaks in Mobile Network Traffic. In *Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services*, MobiSys ’16, pages 361–374, New York, NY, USA, 2016. ACM.

- [152] Ronald W Rogers and Donald L Thistlethwaite. Effects of fear arousal and reassurance on attitude change. *Journal of personality and social psychology*, 15(3):227, 1970.
- [153] Michael Salter. *Crime, justice and social media*. Routledge, 2016.
- [154] Antonella Santi. ” catfishing”: A comparative analysis of us v. canadian catfishing laws & their limitations. *S. Ill. ULJ*, 44:73, 2019.
- [155] Morgan Klaus Scheuerman, Stacy M Branham, and Foad Hamidi. Safe spaces and safe places: Unpacking technology-mediated experiences of safety and harm with transgender people. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–27, 2018.
- [156] Morgan Klaus Scheuerman, Jialun Aaron Jiang, Casey Fiesler, and Jed R Brubaker. A framework of severity for harmful content online. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, 2021.
- [157] Sarita Schoenebeck and Lindsay Blackwell. Reimagining social media governance: Harm, accountability, and repair. *SSRN (July 29, 2021)*, 2021.
- [158] Sarita Schoenebeck, Oliver L Haimson, and Lisa Nakamura. Drawing from justice theories to support targets of online harassment. *new media & society*, page 1461444820913122, 2020.
- [159] Natasha Dow Schüll. The folly of technological solutionism: An interview with evgeny morozov, 2013.
- [160] Mark Scott and Laura Kayali. What happened when humans stopped managing social media content, 2020.
- [161] Nick Seaver. Captivating algorithms: Recommender systems as traps. *Journal of material culture*, 24(4):421–436, 2019.
- [162] Ruth Shillair, Shelia R Cotten, Hsin-Yi Sandy Tsai, Saleem Alhabash, Robert LaRose, and Nora J Rifon. Online safety begins with you and me: Convincing internet users to protect themselves. *Computers in Human Behavior*, 48:199–207, 2015.
- [163] Varyanne Sika and Nanjira Sambuli. Ict4governance in east africa. In *International Conference on e-Infrastructure and e-Services for Developing Countries*, pages 175–179. Springer, 2014.
- [164] Milene Selbach Silveira, Clarisse Sieckenius de Souza, and Simone DJ Barbosa. Semiotic engineering contributions for designing online help systems. In *Proceedings of the 19th annual international conference on Computer documentation*, pages 31–38. ACM, 2001.
- [165] Rashmi Sinha and Kirsten Swearingen. The role of transparency in recommender systems. In *CHI’02 extended abstracts on Human factors in computing systems*, pages 830–831. ACM, 2002.
- [166] Amy K Smith, Ruth N Bolton, and Janet Wagner. A model of customer satisfaction with service encounters involving failure and recovery. *Journal of marketing research*, 36(3):356–372, 1999.
- [167] Troy Smith and Nikolaos Stamatakis. Cyber-victimization trends in trinidad & tobago: The results of an empirical research. *International Journal of Cybersecurity Intelligence & Cyber-crime*, 4(1):46–63, 2021.
- [168] Muhammad Sohaib, Peng Hui, Umair Akram, Abdul Majeed, Zubair Akram, and Muhammad Bilal. Understanding the justice fairness effects on ewom communication in social media environment. *International Journal of Enterprise Information Systems (IJEIS)*, 15(1):69–84, 2019.

- [169] Ashkan Soltani. Abusability testing: Considering the ways your technology might be used for harm. In *Enigma 2019 (Enigma 2019)*, Burlingame, CA, January 2019. USENIX Association.
- [170] Katta Spiel, Oliver L Haimson, and Danielle Lottridge. How to do better with gender on surveys: a guide for hci researchers. *Interactions*, 26(4):62–65, 2019.
- [171] Gaurav Srivastava, Saksham Chitkara, Kevin Ku, Swarup Kumar Sahoo, Matt Fredrikson, Jason Hong, and Yuvraj Agarwal. PrivacyProxy: Leveraging Crowdsourcing and In Situ Traffic Analysis to Detect and Mitigate Information Leakage. *arXiv preprint arXiv:1708.06384*, 2017.
- [172] Stefan Stieglitz and Linh Dang-Xuan. Emotions and information diffusion in social media—sentiment of microblogs and sharing behavior. *Journal of management information systems*, 29(4):217–248, 2013.
- [173] Angelika Strohmayer, Julia Slupska, Rosanna Bellini, Lynne Coventry, Tara Hairston, and Adam Dodge. Trust and abusability toolkit: Centering safety in human-data interactions. 2021.
- [174] Christian Sturm, Alice Oh, Sebastian Linxen, Jose Abdelnour Nocera, Susan Dray, and Katharina Reinecke. How weird is hci? extending hci principles to other countries and cultures. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, pages 2425–2428, 2015.
- [175] Cass R Sunstein. Algorithms, correcting biases. *Social Research: An International Quarterly*, 86(2):499–511, 2019.
- [176] P. Symeonidis, A. Nanopoulos, and Y. Manolopoulos. Providing justifications in recommender systems. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 38(6):1262–1272, 2008.
- [177] Omer Tene and Jules Polonetsky. Taming the golem: Challenges of ethical algorithmic decision-making. *NCJL & Tech.*, 19:125, 2017.
- [178] Dhanaraj Thakur. How do icts mediate gender-based violence in jamaica? *Gender & Development*, 26(2):267–282, 2018.
- [179] Nava Tintarev and Judith Masthoff. Evaluating the effectiveness of explanations for recommender systems. *User Modeling and User-Adapted Interaction*, 22(4):399–439, issn=, Oct 2012.
- [180] Rina Torchinsky. How period tracking apps and data privacy fit into a post-roe v. wade climate. *NPR*, 2022.
- [181] Sabine Trepte, Leonard Reinecke, Nicole B Ellison, Oliver Quiring, Mike Z Yao, and Marc Ziegele. A cross-cultural perspective on the privacy calculus. *Social Media+ Society*, 3(1):2056305116688035, 2017.
- [182] Blase Ur and Yang Wang. A cross-cultural framework for protecting user privacy in online social media. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 755–762, 2013.
- [183] José Van Dijck. Facebook as a tool for producing sociality and connectivity. *Television & new media*, 13(2):160–176, 2012.
- [184] Viswanath Venkatesh, Susan A Brown, Likoebe M Maruping, and Hillol Bala. Predicting different conceptualizations of system use: The competing roles of behavioral intention, facilitating conditions, and behavioral expectation. *MIS quarterly*, pages 483–502, 2008.



- [185] Jin Ho Verdonshot, Maurits Barendrecht, Laura Klaming, and Peter Kamminga. Measuring access to justice: The quality of outcomes. *Tilburg University Legal Studies Working Paper*, (014), 2008.
- [186] Andrea C Villanti, Amanda L Johnson, Vinu Ilakkuvan, Megan A Jacobs, Amanda L Graham, and Jessica M Rath. Social media use and access to digital technology in us young adults in 2016. *Journal of medical Internet research*, 19(6):e7303, 2017.
- [187] Emily A Vogels. The state of online harassment. *Pew Research Center*, 13, 2021.
- [188] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Why fairness cannot be automated: Bridging the gap between eu non-discrimination law and ai. *Computer Law & Security Review*, 41:105567, 2021.
- [189] Todd Wasserman. Facebook to hit 1 billion user mark in august. *Last retrieved on 12th of April*, 2012.
- [190] Maranke Wieringa. What to account for when accounting for algorithms: A systematic literature review on algorithmic accountability. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, FAT\* '20*, page 1–18, New York, NY, USA, 2020. Association for Computing Machinery.
- [191] Darcia Wilkinson, Öznur Alkan, Q Vera Liao, Massimiliano Mattetti, Inge Vejsbjerg, Bart P Knijnenburg, and Elizabeth Daly. Why or why not? the effect of justification styles on chatbot recommendations. *ACM Transactions on Information Systems (TOIS)*, 39(4):1–21, 2021.
- [192] Darcia Wilkinson, Paritosh Bahirat, Moses Namara, Jing Lyu, Arwa Alsubhi, Jessica Qiu, Pamela Wisniewski, and Bart P Knijnenburg. Privacy at a glance: the user-centric design of glanceable data exposure visualizations. *Proceedings on Privacy Enhancing Technologies*, 2020(2):416–435, 2020.
- [193] Darcia Wilkinson and Bart P. Knijnenburg. Many Islands, Many Problems: An Empirical Examination of Online Safety Behaviors in the Caribbean. In *The Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (Under Review)*, pages 1–5, 2022.
- [194] Darcia Wilkinson, Moses Namara, Karla Badillo-Urquiola, Pamela J Wisniewski, Bart P Knijnenburg, Xinru Page, Eran Toch, and Jen Romano-Bergstrom. Moving beyond a” one-size fits all” exploring individual differences in privacy. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–8, 2018.
- [195] Darcia Wilkinson, Moses Namara, Karishma Patil, Lijie Guo, Apoorva Manda, and Bart Knijnenburg. The pursuit of transparency and control: A classification of ad explanations in social media. 2021.
- [196] Pamela Wisniewski, Haiyan Jia, Na Wang, Saijing Zheng, Heng Xu, Mary Beth Rosson, and John M Carroll. Resilience mitigates the negative effects of adolescent internet addiction and online risk exposure. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 4029–4038, 2015.
- [197] Irene Woon, Gek-Woo Tan, and R Low. A protection motivation theory approach to home wireless security. 2005.
- [198] Liang Wu, Fred Morstatter, Kathleen M Carley, and Huan Liu. Misinformation in social media: definition, manipulation, and detection. *ACM SIGKDD Explorations Newsletter*, 21(2):80–90, 2019.

- [199] Sijia Xiao, Coye Cheshire, and Niloufar Salehi. Sensemaking, support, safety, retribution, transformation: A restorative justice approach to understanding adolescents’ needs for addressing online harm. In *CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2022.
- [200] Heng Xu, Sumeet Gupta, Mary Beth Rosson, and John M Carroll. Measuring mobile users’ concerns for information privacy. 2012.
- [201] Heng Xu, Hock-Hai Teo, Bernard CY Tan, and Ritu Agarwal. Research note-effects of individual self-protection, industry self-regulation, and government regulation on privacy concerns: a study of location-based services. *Information Systems Research*, 23(4):1342–1363, 2012.
- [202] Crystal S Yang and Will Dobbie. Equal protection under algorithms: A new statistical and legal framework. *Michigan Law Review*, 119(2):291–395, 2020.
- [203] Jillian C York. *Silicon Values: The Future of Free Speech Under Surveillance Capitalism*. Verso Books, 2021.
- [204] Seounmi Youn. Teenagers’ perceptions of online privacy and coping behaviors: a risk–benefit appraisal approach. *Journal of Broadcasting & Electronic Media*, 49(1):86–110, 2005.
- [205] Douglas Zytke, Pamela J. Wisniewski, Shion Guha, Eric PS Baumer, and Min Kyung Lee. Participatory design of ai systems: Opportunities and challenges across diverse users, relationships, and application domains. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, pages 1–4, 2022.