Many Islands, Many Problems: An Empirical Examination of Online Safety Behaviors in the Caribbean

Daricia Wilkinson
dariciw@clemson.edu
Clemson University
Clemson, SC, USA

Bart P. Knijnenburg
bartk@clemson.edu
Clemson University
Clemson, SC, USA

ABSTRACT

Little is known about non-Western social media users' motivations for adopting behaviors that protect them against pervasive threats to their privacy, security, and personal well-being. Drawing on Rogers' Protection Motivation Theory (PMT), this survey study explores Caribbean people's (N=551) perceptions of safety threats and the factors contributing to their intention to adopt protective behaviors. Our analysis revealed that prior victimization was associated with increased perceptions of vulnerability and severity of harms, which, in turn, influenced elevated safety protection behaviors. For harassment-related harms in particular, participants' trust in social media sites increased their intention to adopt protective behaviors. We observe significant country-to-country differences, which we contextualize through interviews with experts throughout the region. Our findings provide a new understanding of users' mental models, behaviors, and attitudes with respect to online safety. We conclude by discussing theoretical and practical implications and outline opportunities for the design of inclusive and culturally-aware safety tools.

CCS CONCEPTS

• **Human-centered computing → Human computer interaction (HCI)**; • **Security and privacy → Human and societal aspects of security and privacy**.

KEYWORDS

Online safety, social media, protection motivation theory, Caribbean, survey

ACM Reference Format:
Daricia Wilkinson and Bart P. Knijnenburg. 2022. Many Islands, Many Problems: An Empirical Examination of Online Safety Behaviors in the Caribbean. In *CHI Conference on Human Factors in Computing Systems (CHI '22), April 29-May 5, 2022, New Orleans, LA, USA.* ACM, New York, NY, USA, 25 pages. https://doi.org/10.1145/3491102.3517643

1 INTRODUCTION

In 2017, Jason Jones successfully filed a lawsuit against the government of Trinidad and Tobago claiming that the sexual offenses act [1] criminalized intimacy between consenting same-sex adults and infringed on the rights of LGBTQ+ citizens [18]. Shortly after filing the law suit, Jones received over 50 death threats on social media by persons wanting to burn him alive. He hired a security guard and moved from his apartment to protect his roommate. During this time, he expressed the emotional toll of the online harassment, especially in light of little amount of support he received from his family [63]. The incidents arising from this experience illustrate that there are multiple harms that arise from digital interactions that jeopardize people's physical, emotional, and relational safety in both online and offline contexts.

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 [3, 4, 13, 31, 61]. 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) [50].

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=551), 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) [36], to understand what factors motivate Caribbean social

[1] Sexual offences act of Trinidad and Tobago : http://rgd.legalaffairs.gov.tt/laws2/alphabetical_list/lawspdfs/11.28.pdf

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
CHI '22, April 29-May 5, 2022, New Orleans, LA, USA
© 2022 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-9157-3/22/04…$15.00
https://doi.org/10.1145/3491102.3517643

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?

In addressing each of the proposed research questions, we consider sample-wide trends to understand the landscape throughout the region as well as country-to-country differences to acknowledge localized needs. Our results offer support for the three stage logic of PMT—prior victimization influences how people evaluate threats, which in turn impacts their intention to adopt safety strategies. We also found that there are nuances in safety intentions for online versus offline threats, which has implications for policy and design. Additionally, although there is a shared socio-historical culture throughout the region, our analysis revealed country-to-country differences in perceptions of severity, vulnerability, and intentions to engage in protective behaviors.

By employing structural equation modeling, we were able to provide empirical evidence on the relationship between protection motivation and safety intentions online. Specifically, our paper makes the following research 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.

In the following sections, we discuss literature related to our work and provide theoretical support for the framing of the concepts that ground the work. We then describe our methodology, followed by a presentation of the results and a subsequent discussion of the implications of the work. We conclude with opportunities for future work.

2 BACKGROUND AND THEORETICAL DEVELOPMENT

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. Lastly, we present our hypotheses for the study.

2.1 Online Threats and Safety Protection

2.1.1 Adopting a Wider Lens on Online Safety. Inherently, the design of social networking systems encourages online interactions, which has proven to have immense benefits for discourse, social support, and overall well-being [3, 4, 13, 31, 61]. 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 [10, 49]. 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 [25]. These evolving threats and vulnerabilities require an expansion of our understanding of these 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 have argued it is now a human rights violation [43]. In a similar light, misinformation online has influenced elections and highlighted its potential as a viable threat to democracy [67].

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 [47]. 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 [50]. 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. [44], 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 depart from the approach to development focused on building fast and fixing later [56]. Soltani argued that building safer technology requires a comprehensive testing of platforms' vulnerability to being abused and that teams need to adopt abusability testing [55]. 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 [25, 30, 60]. However, much of the work that focuses on protective behaviors has (1) largely been focused on Western, Educated, Industrialized, Rich, and Democratic (WEIRD) cultures [32, 57], 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, 25, 41]. 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 [57]. 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 [25]. Our work complements and expands on this work by considering 15 countries across a region often excluded in HCI research.

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, there are wide variations exist. Even though they are geographically closely located, each country has unique attributes and challenges. 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 [25, 47, 52] including within the Caribbean region [12]. 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 [14]. Caribbean leaders pushed for the creation of the Caribbean Community Implementation Agency for Crime and Security (CARICOM IMPACS)[2] 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 [42]. 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 in 1988 in developing GDPR" [35].

[2]CARICOM IMPACS: https://caricomimpacs.org/cyber-security/

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 [58]. 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 [54]. 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.

2.2 Protection Motivation Theory

Protection Motivation Theory (PMT) [36, 48] 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 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 [36, 48].

Recent studies that have applied PMT in the context of online safety have investigated the motivation behind using computer virus protection [29], online privacy [69], harassment [34], predicting internet scam victimization [16] and digital security [53]. 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

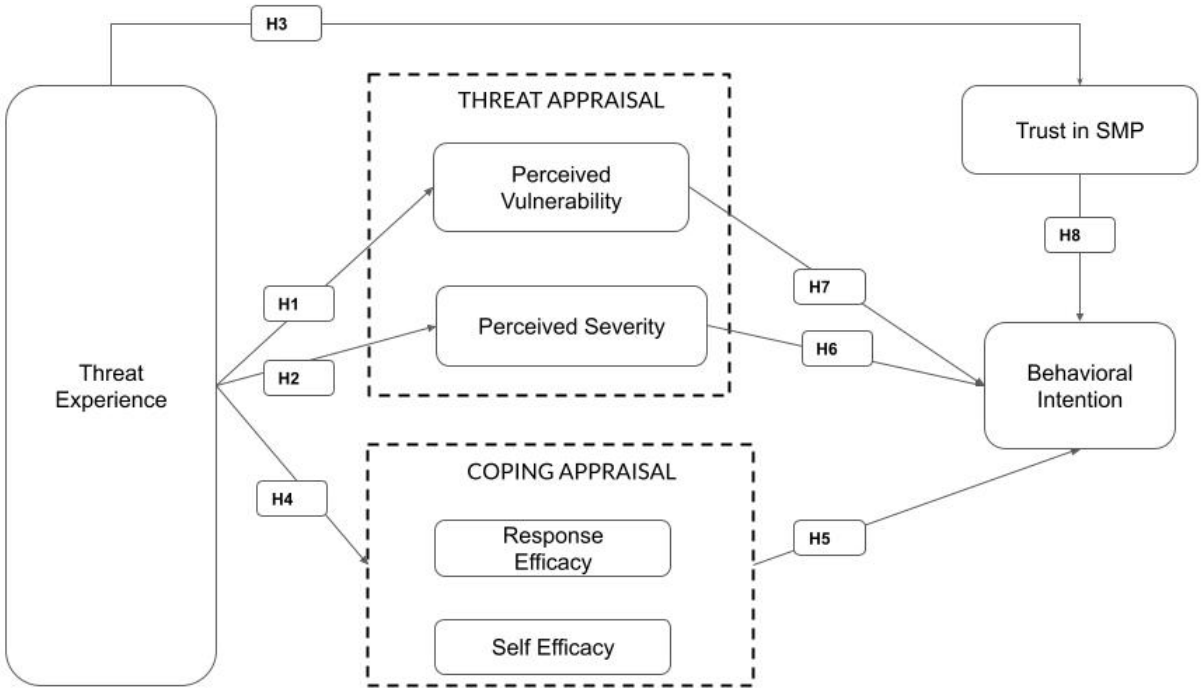

Figure 1: The figure above illustrates the proposed conceptual model for the study

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.

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 [20]. 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 [17]. This likely happens because those who have personally been victims of safety harms are likely to understand the severe consequences associated with that threat [8].

For example, Mohamed and Ahmad found that persons who were victims of internet scams tended to build more knowledge about related severity and vulnerability [39]. 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) [20]. 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 [48]. Woon found that increased levels of perceived severity positively affected participants' security behavior [66]. 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 [26]. 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

[34]. 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 [36]. 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 [38]. Studies have shown that social media users are more likely to trust platforms that could keep them protected from safety harms [2]. Kim et al. illustrated how usable privacy policies predicted consumers' trust of a website [27]. Conversely, social media companies have faced increasing public pressure because of risks to users' safety, such as unfair data collection [68], harassment [9, 64], and overall concerns for better safety tools [47]. Based on these findings, we hypothesize the following:

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

3 OVERVIEW OF METHODS

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.

3.1 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 1.

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, 12 responses were excluded from the analysis due to low quality or a low number of responses for that country, which left a total sample size of 551.

3.2 Participants

We treated age, gender, ethnicity, household income, education, and sexual orientation as exogenous variables. The majority of respondents identified as Black or Afro-Caribbean (70.24%), followed by Kalinago (3.25%)[3], White or Caucasian (2.11%), East Indian, (3.09%), Asian (0.81%), American Indian or Alaska Native (0.81%), Hispanic (0.81%), two or more races (5.85%), not disclosed (6.83%). The average age of respondents was 29 years (SD = 9.31 years). Household income was measured on a 13 point scale from "less than $4,999 USD" to "$65,000 USD or more". Around 62% of respondents reported a household income below $34,999 annually. Education was measured on a 8 point scale ranging from "less than a high school degree" to "professional degree (e.g. JD or MD)". Education levels varied: 7.10% had less than a high school degree, 42.83% obtained medium levels of education (secondary education), and 41.55% obtained higher education (bachelor's and beyond). The sample consisted of 59% women. Regarding sexual orientation, 76% identified as heterosexual, bisexual (8%), gay (2%), self-described (2%), and the remainder preferred not to disclose.

3.3 Measurement

Prior work has measured threats to online safety in multiple ways, in both technical and non-technical contexts [47, 54]. Where possible, we adopted validated measurement scales, while some scales were adapted from existing work that focused on similar issues. Given the complexity of online threats, we felt that it was crucial to capture a wide range of experiences, as different types of threats co-occur on large scales on social media. Recent research has acknowledged the entangled nature of threats and the importance of studying a variety of risks to better understand how they relate to each other [47, 50]. Based on this notion, we considered both technical and non-technical risks and their respective protective behaviors. We chose to ask participants questions about their intentions to adopt protective behaviors, as intentions are acknowledged as one of the most significant predictors of actual behaviour [19]. Furthermore, our methods are aligned with similar studies that explored similar online safety contexts and measured intentions based on a range of behaviors [16, 24, 26, 34, 53, 54].

We consider a wide range of threats, including threats related to:

- Digital security: Instances that risk the protection of a person's personal account and/or files from intrusion by an outside user.
- Access and disclosure: Threats that risk information privacy and occur as a result of unacceptable or unwanted data collection, processing, or sharing.
- Harassment: Interpersonal interacts that potentially harm or damage an individual or impacts them negatively.
- Online to Offline: Interactions that begin in digital contexts but has post-digital consequences spilling into offline contexts.

[3]Persons identifying as Kalinago are members of an indigenous tribe of people in the Lesser Antilles in the Caribbean.

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 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.

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 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*.

Iterative feedback was gathered from a group of 15 Human Computer Interaction (HCI) researchers over the course of two months to ensure the usability of the survey. Two local research assistants were also recruited from the Caribbean to ensure cultural relevance of the measurement instrument. Upon extensive discussions, we decided to make explicit distinctions between offline and online stalking, as well as adding items such as "your phone was cloned by someone without permission". These were items that local researchers expressed were prevalent regionally and important to have included. Aside from Threat Experience (which was measured with dichotomous items), all remaining factors were measured on 7-point Likert-type scales to ensure uniformity and comparability (see Tables 5-8 in the Appendix for the full lists of items).

3.3.1 Threat Experience. Similar to Chen et al. [15], threat experiences were measured with dichotomous items (yes or no), asking respondents whether they had experienced any harms (See Figure 2 for breakdown of reported threat experiences). This aligns with similar research that applied PMT to predict protection behavior [34].

3.3.2 Threat appraisal. The perceived severity of a threat was assessed by asking respondents to indicate how serious a particular threat (e.g., "your login information being at risk of being compromised") was to them, while their perceived vulnerability to that threat was assessed by asking respondents about how likely they think they would experience each of the threats.

3.3.3 Coping appraisal. To measure self efficacy, respondents indicated the extent to which they believed the could employ particular protective behaviors (e.g., "set up Login alert for my social media accounts"). For response efficacy, participants were asked to what extent they agreed that the respective protective behaviors helped them feel safe online. Protective behaviors were chosen that mirrored the threats they were designed to address (see Tables 5-8 in the Appendix for the alignment between threats and protective behaviors).

3.3.4 Behavioral Intention. Behavioral intention has been tested extensively as a reliable construct for predicting human behavior [62]. For this construct, we asked respondents about their intention to adopt particular protective behaviors in the event of an unsafe online experience.

3.3.5 Other Scales. In addition to the constructs mentioned above, we also explored respondents' *social media usage* using a scale adopted from Perrin and Anderson [45]. Furthermore, five items were adopted from the Internet users' information privacy concerns (IUIPC) scale to measure respondents' *trust in social media platforms* [37]. In addition, we adapted Lusoli et al.'s scale measuring perceptions of *third party responsibility* for the safety context [33].

3.4 Positionality Statement

This project comprises of co-authors who are primarily associated with one US-based university. However, the work was carefully conducted to examine the underlying challenges facing the Caribbean community while ensuring local experts were leading research efforts. To align with our goal of de-centering Western perspectives, we chose to include these experts in each phase of the study to ensure cultural relevance and that perspectives from different islands were included. In addition, we worked alongside a non-profit organization in the Caribbean to conduct the work. As a result, the team for this project consisted of a graduate student, faculty, local collaborators and local research assistants. Many of the project members were physically present in the region throughout the duration of the study. At the conclusion of the survey data collection, we once again involved local experts to position the findings within the context of the region. Their input helped to formulate and situate the implications of this work.

4 RESULTS

We organize our results by the initial research questions outlined in section 2.3 and present the findings related to our proposed conceptual model.

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 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 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.

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 protective mechanisms [25], 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 4 and Figure 5 for a breakdown across different threats).

It can be referred from data in Figure 5 that, compared to the severity levels displayed in Figure 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 it is a major threat, most participants were not convinced they were likely to have that experience.

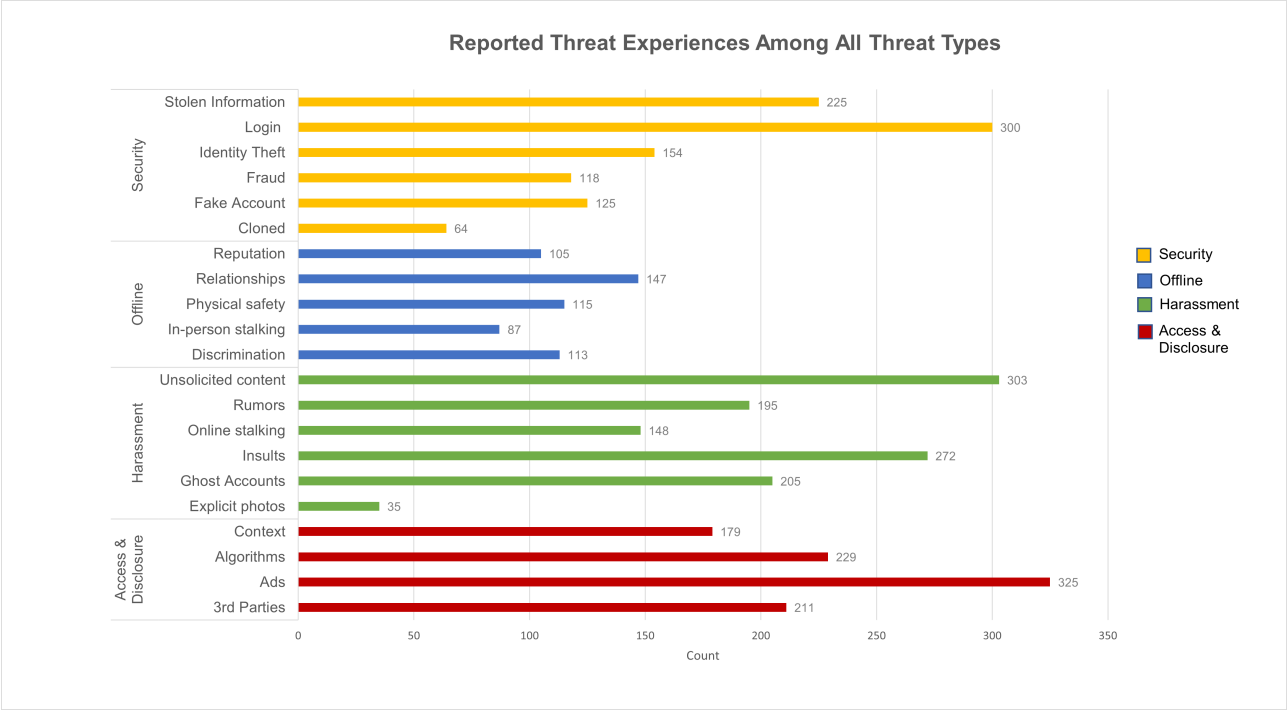

Figure 2: Reported prior victimization counts across all participants (N=551) and all observed threat categories.

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: 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)

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 [50]. For each factor, we use multi-item measurement scales to control for measurement error [23].

To validate the robustness and validity of our measurement scales, Confirmatory Factor Analysis (CFA) was employed. Items

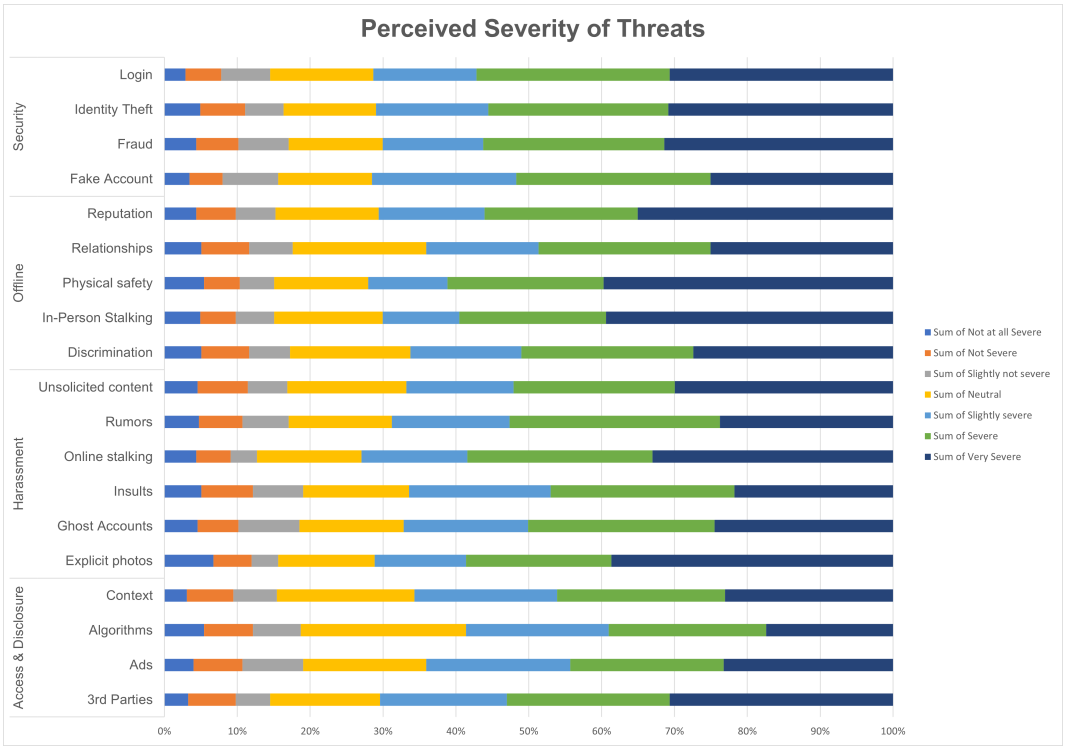

Figure 4: Sample-wide comparison of the perceived severity of threats across all threat categories

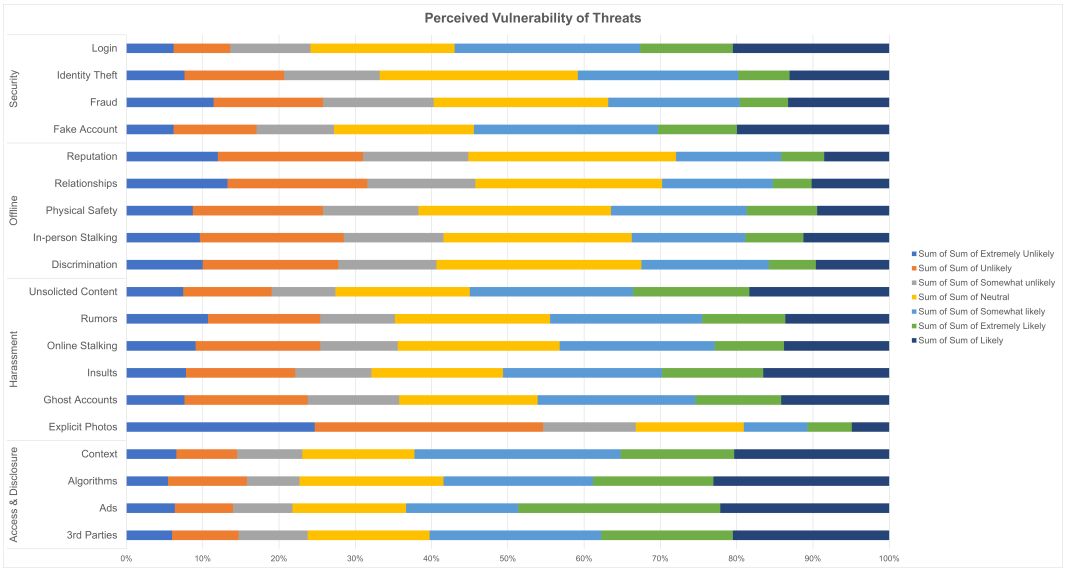

Figure 5: Sample-wide comparison of the perceived vulnerability of threats across all threat categories

with low loadings were removed from subsequent analyses (see the greyed-out items in Tables 5-8 in the Appendix).

Discriminant validity was assessed by comparing the average variance extracted (AVE) of 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 α values were excellent[4], 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[5] with the evaluation results are presented in Figures 6-9. The model fit indices for all four models indicate good to excellent fit[6].

- Threats related to online-to-offline contexts: excellent fit: $\chi^2(315)= 608.795$, $p < .01$; RMSEA = 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. For clarity, we report significant direct effects from left to right and endogenous variable are not depicted.

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 ($\beta = 0.571$, $p < 0.001$), digital security ($\beta = 0.390$, $p < 0.001$), access & disclosure ($\beta = 0.672$, $p < 0.001$), and online-to-offline contexts ($\beta = 0.583$, $p < 0.001$). Therefore, H1 is supported.

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 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 13, 10, 11, 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.

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 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 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. [7]). 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 behaviors for threats related to their digital security ($\beta = 0.162$, $p < 0.01$) and access and disclosure of their personal data ($\beta = 0.152$, $p < 0.01$). Thus, H6 is only supported for

[4]For alpha, >.70 is acceptable, >.80 is good, >.90 is excellent.
[5]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 β coefficients (and the standard error) of the effect
[6]A model should not have a non-significant χ^2, but this statistic is regarded as too sensitive [6]. Hu and Bentler [22] 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.

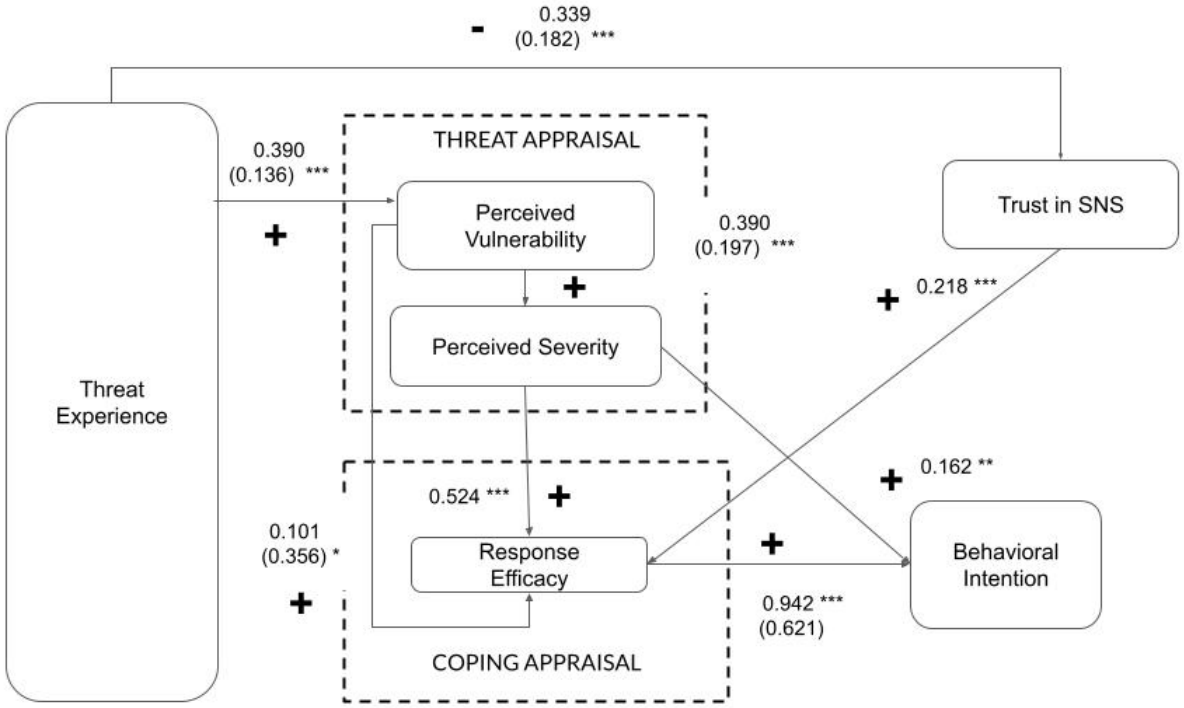

Figure 6: The figure above displays the SEM models for threats related to digital security

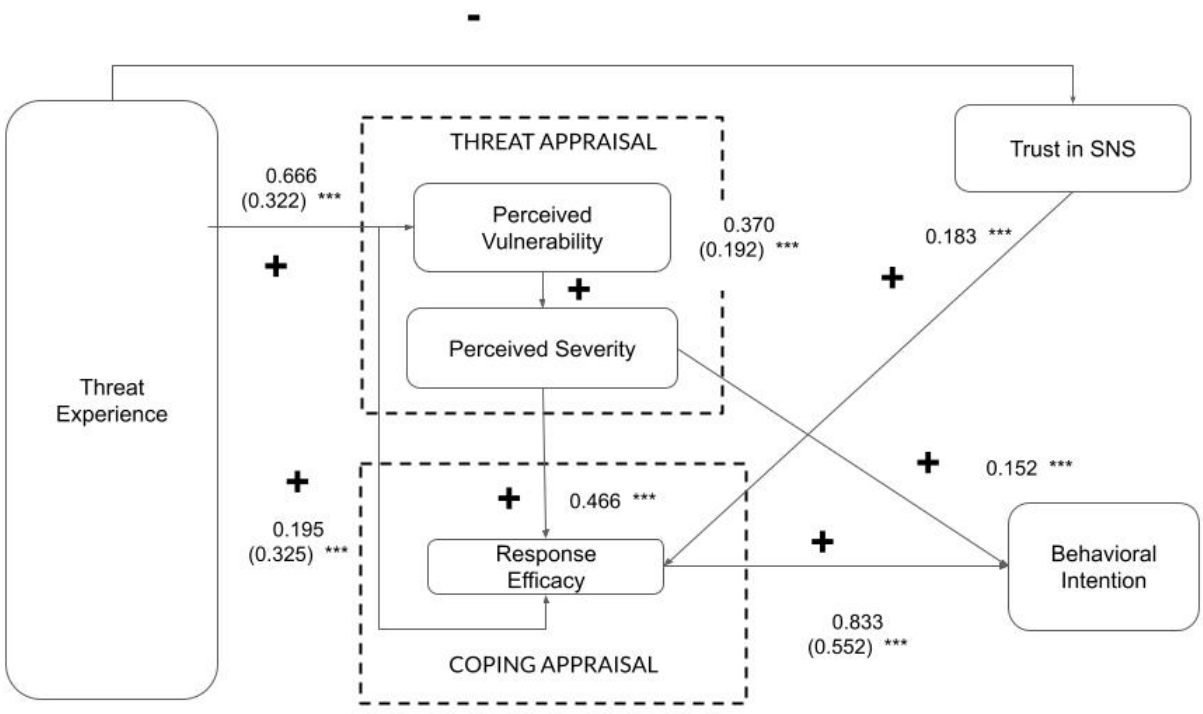

Figure 7: The figure above displays the SEM models for threats related to the access and disclosure of personal information.

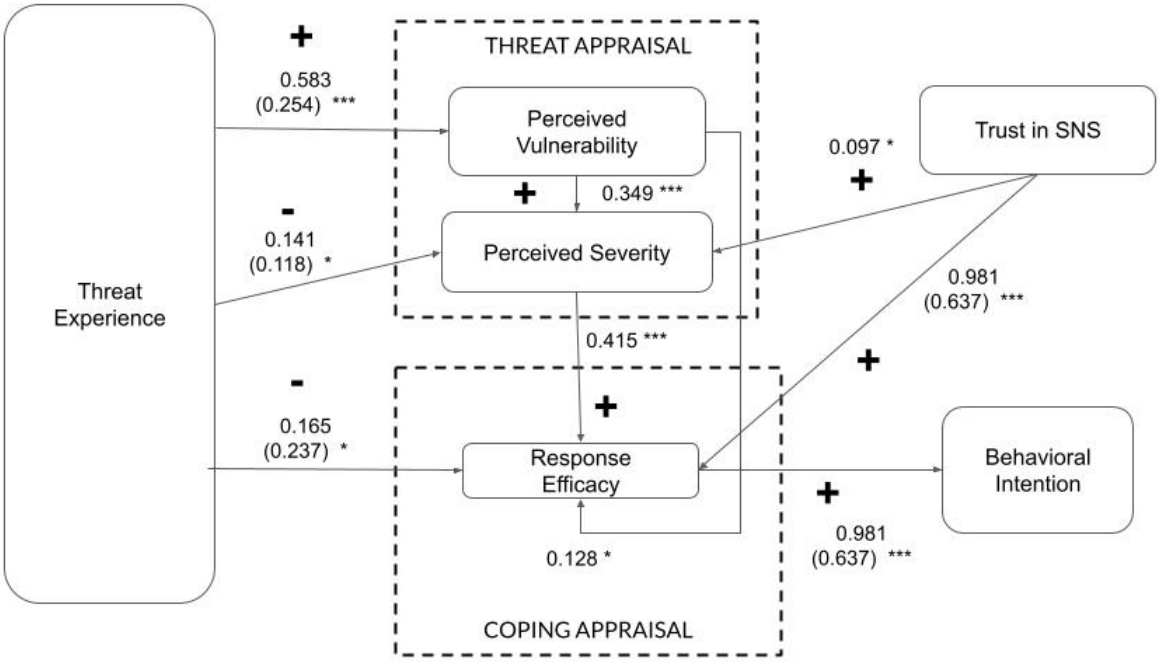

Figure 8: The figure above displays the SEM models for threats related to online-offline contexts

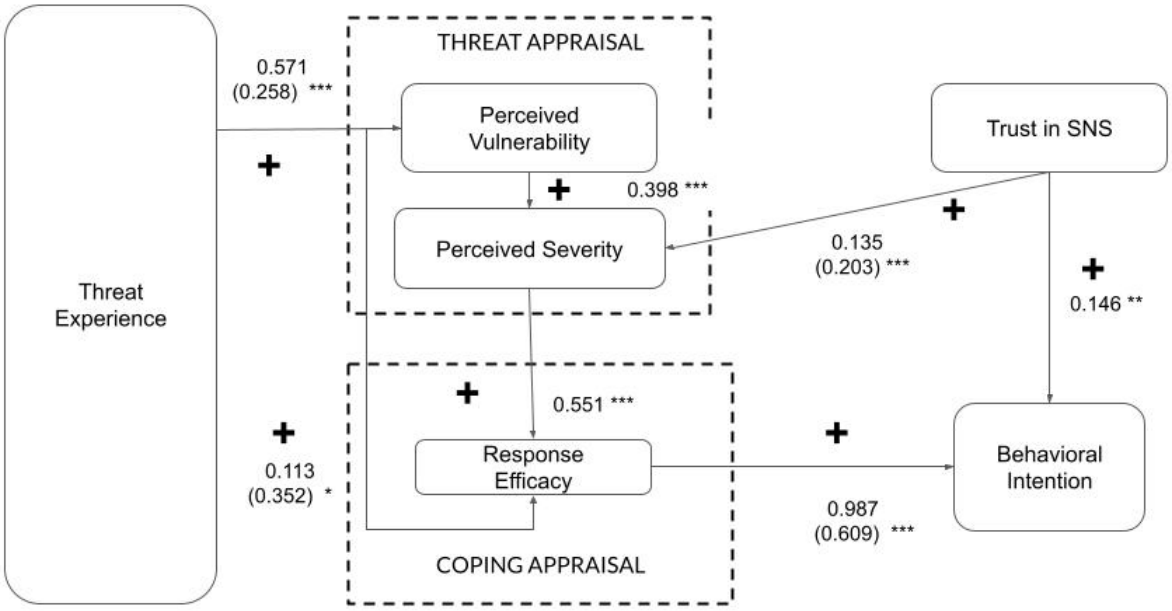

Figure 9: The figure above displays the SEM models for Harassment-related threats

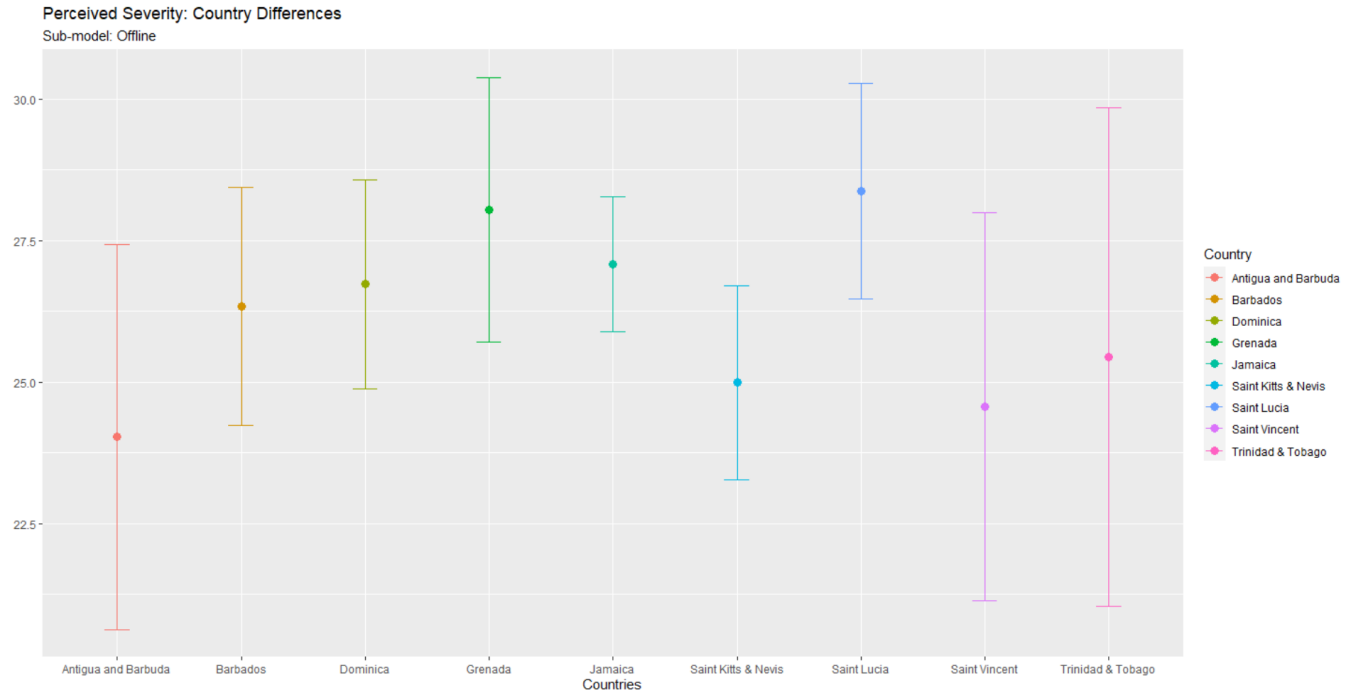

Figure 10: Marginal effects of perceived severity for online-to-offline threats

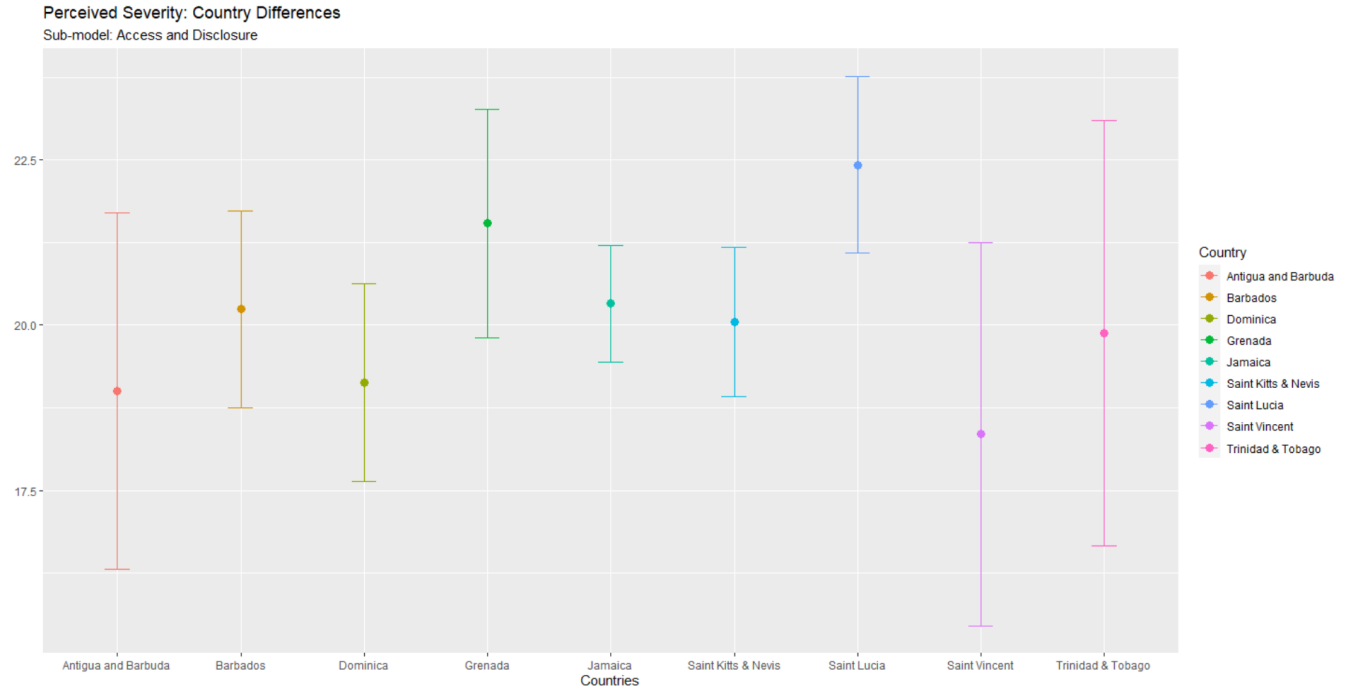

Figure 11: Marginal effects of perceived severity for threats related to Access and Disclosure

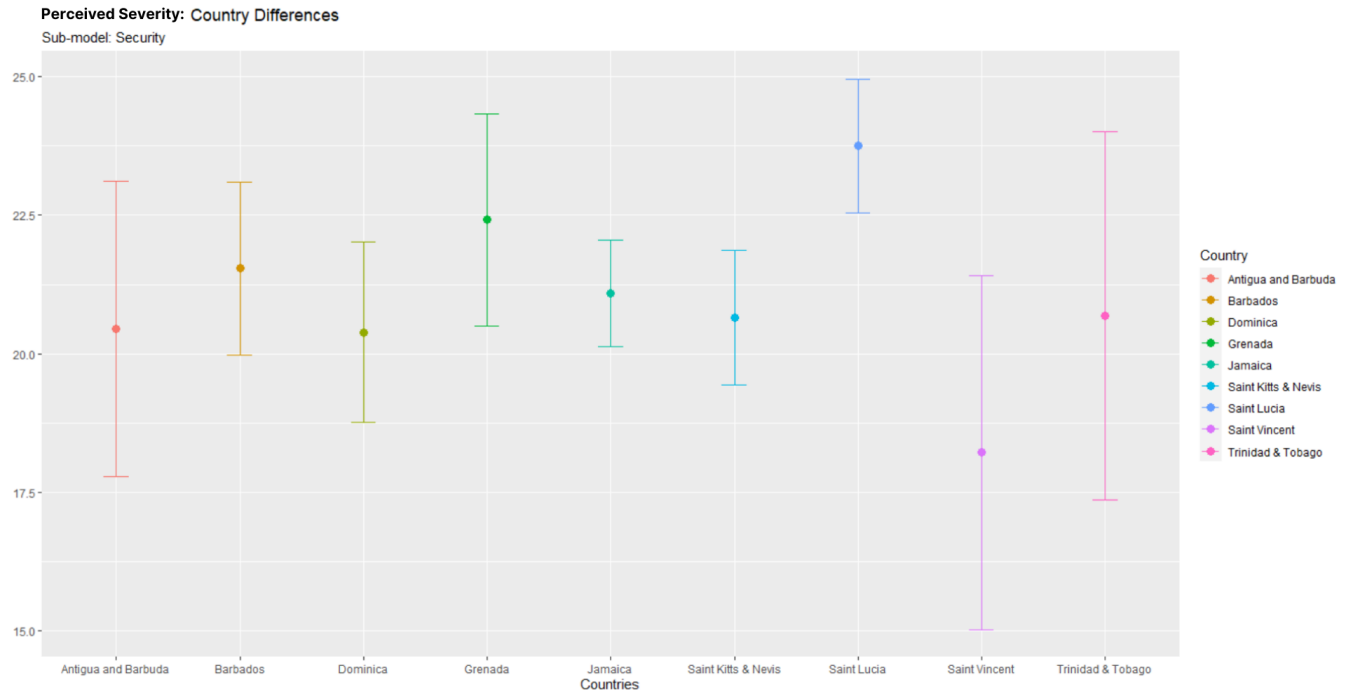

Figure 12: Marginal effects of perceived severity for security threats

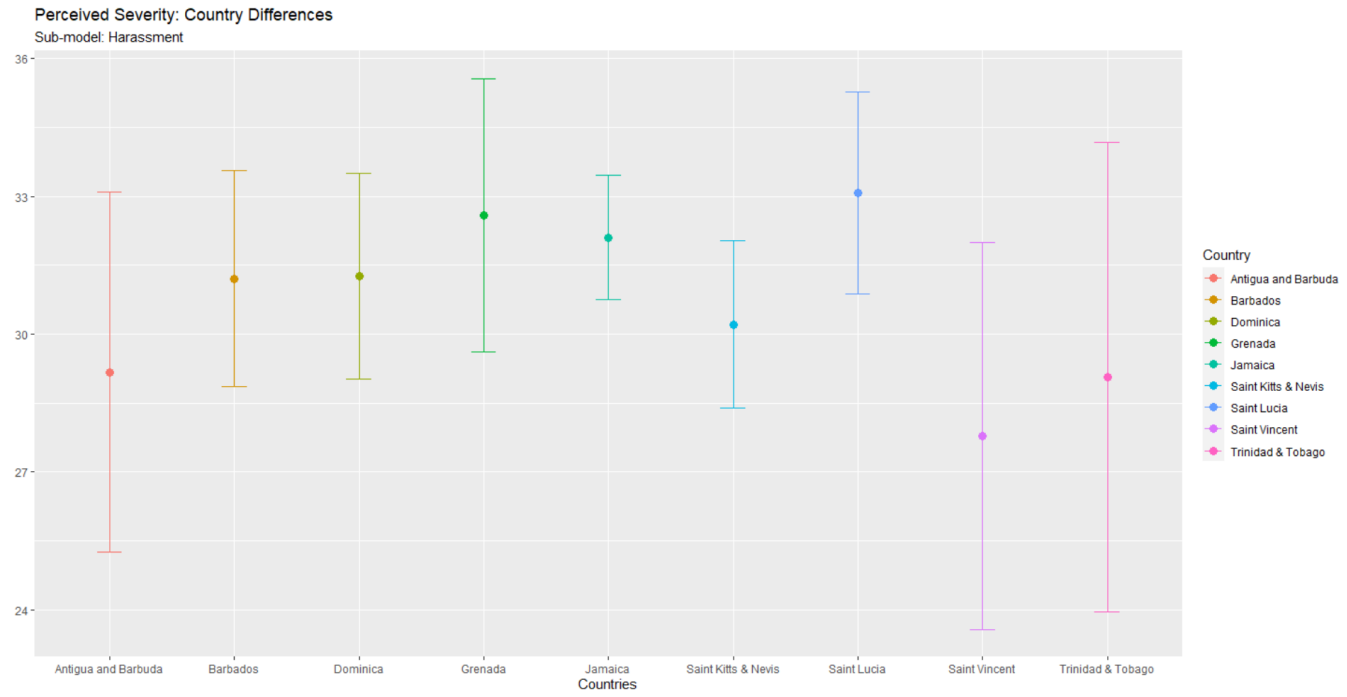

Figure 13: Marginal effects of perceived severity for harassment-related threats

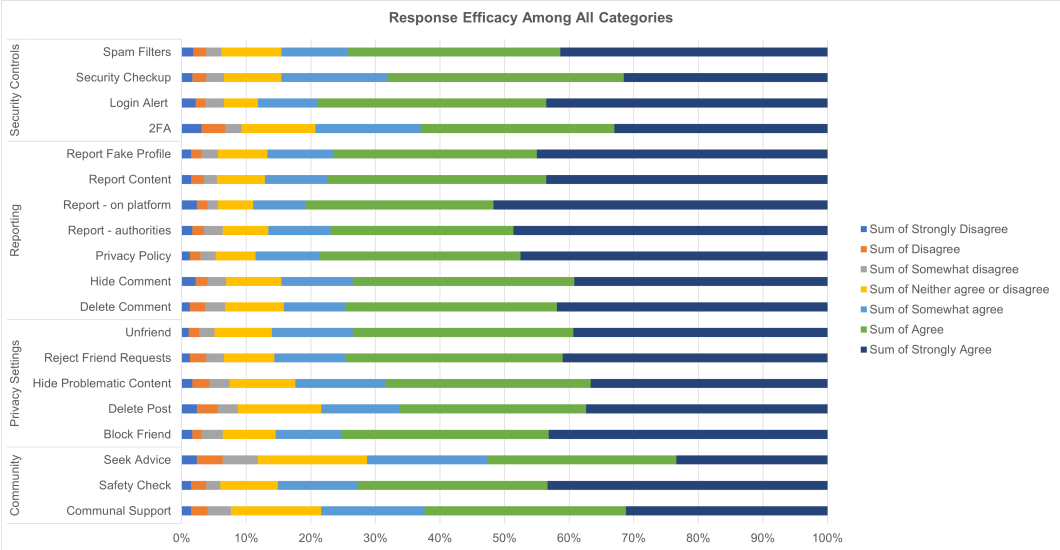

Figure 14: Sample-wide comparison of the response efficacy across all protective behavior categories

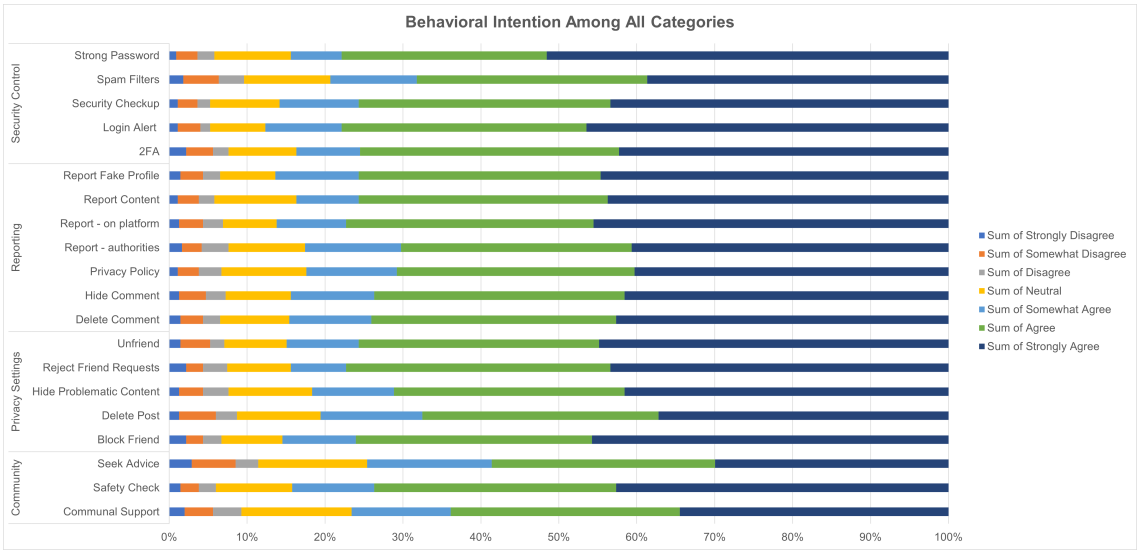

Figure 15: Sample-wide comparison of behavioral intention across all protective behavior categories

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 ($\beta = 0.146$, $p < 0.01$). 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 users increase their response efficacy against them.

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: 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.

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 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.

We summarize the findings in tables 3 and 4.

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.

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 challenging 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, 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 [36], 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, and higher vulnerability resulted in higher perceived severity, which in turn increased users' response efficacy, which then in turn increased their intentions to engage with protective behaviors. Researchers have argued that risk exposure builds resilience and aids in risk mitigation [65]. 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—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 may 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, one of 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 [5, 21, 40]. 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 [46]. 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 [28, 30]. 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) [59]. 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.

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 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 [42]. 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 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 [11]. 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 [51]. 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 to design and develop culturally-aware online safety tools.

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 awareness 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.

5.3 Limitations and Future Work

In our study, we chose to investigate a wide variety of different types of harms that might be sensitive and trigger negative past experiences. As a result, persons might have felt embarrassed revealing their experiences in answering our survey. Additionally, the threats represented in our survey are not exhaustive, and there might be other threats that Caribbean people experience that are not captured. While it would be challenging to identify all possible threats, future studies can examine additional threats to safety and how these impact special populations such as our youth.

Secondly, we capture users' behavioral intention to adopt protective measures. Although studies have shown that this is a reliable determinant of actual behavior, future work can conduct longitudinal studies to explore the relationship between previous exposure to harms and users' actual usage of protective behaviors.

Finally, it should be noted that the recruitment of participants for our study required additional considerations since many popular crowd sourcing platforms typically used in HCI research (such as Amazon Mechanical Turk or Prolific to name a few) do not include respondents from this area. Therefore, reaching some countries proved to be challenging from many aspects. For example, collecting a sample from Haiti turned out to be impossible within the time frame and budget of our study, since this country in particular faced political unrest and natural disasters during the time of data collection. Even in better times, collecting data in Haiti would have required hiring a local company to deploy the study in-person rather than online given limitations in internet and technology access. Generally speaking, it was important that we employed a combination of methods to reach a broad sample of participants. Going forward, researchers could consider studying additional countries in the region, including non-English speaking islands such as Cuba, Puerto Rico, Haiti, and the Dominican Republic. Since there are no easily accessible research panels already in place, future scholars may consider atypical incentives such as offering mobile credit to appeal to a broad sample. This proved to be advantageous in deploying throughout most CARICOM territories but it may be more challenging, for example, in US-based territories such as Puerto Rico or the USVI where that structure is not in place. We note that our sample was skewed towards female participants which may or may not impact the generalizability of our results. Recruiting a balanced sample is often challenging and that difficulty is multiplied when recruiting in over 15 countries. The snowball recruiting method has its limitations but it was proven to be more appropriate and feasible for the study.

6 CONCLUSION

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. We found 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. Our findings shed light on opportunities for both design and policy for the HCI community.

REFERENCES

[1] Rediet Abebe, Kehinde Aruleba, Abeba Birhane, Sara Kingsley, George Obaido, Sekou L Remy, and Swathi Sadagopan. 2021. Narratives and counternarratives on data sharing in Africa. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 329–341.

[2] Alessandro Acquisti and Ralph Gross. 2006. Imagined communities: Awareness, information sharing, and privacy on the Facebook. In *International workshop on privacy enhancing technologies*. Springer, 36–58.

[3] Nazanin Andalibi, Oliver L Haimson, Munmun De Choudhury, and Andrea Forte. 2016. 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*. 3906–3918.

[4] Nazanin Andalibi, Pinar Ozturk, and Andrea Forte. 2017. 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*. 1485–1500.

[5] Antigua News Room. 2019. Husband fined for posting nude photos of wife on social media. https://antiguanewsroom.com/barbados-husband-fined-for-posting-nude-photos-of-wife-on-social-media/

[6] Peter M Bentler and Douglas G Bonett. 1980. Significance tests and goodness of fit in the analysis of covariance structures. *Psychological bulletin* 88, 3 (1980), 588.

[7] Morvareed Bidgoli, Bart P. Knijnenburg, and Jens Grossklags. 2016. When cybercrimes strike undergraduates. In *2016 APWG Symposium on Electronic Crime Research (eCrime)*. 1–10. https://doi.org/10.1109/ECRIME.2016.7487948 ISSN: 2159-1245.

[8] Morvareed Bidgoli, Bart P. Knijnenburg, Jens Grossklags, and Brad Wardman. 2019. Report Now. Report Effectively. Conceptualizing the Industry Practice for Cybercrime Reporting. In *2019 APWG Symposium on Electronic Crime Research (eCrime)*. 1–10. https://doi.org/10.1109/eCrime47957.2019.9037577 ISSN: 2159-1245.

[9] Lindsay Blackwell, Jill Dimond, Sarita Schoenebeck, and Cliff Lampe. 2017. Classification and its consequences for online harassment: Design insights from heartmob. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 1–19.

[10] Lindsay Blackwell, Mark Handel, Sarah T Roberts, Amy Bruckman, and Kimberly Voll. 2018. Understanding" Bad Actors" Online. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–7.

[11] Amelia Bleeker. 2020. 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).

[12] U.S. Embassy Bridgetown. 2021. Over Ninety Media Professionals in the Eastern Caribbean benefit from Media and the Law Training. https://bb.usembassy.gov/over-ninety-media-professionals-in-the-eastern-caribbean-benefit-from-media-and-the-law-training/

[13] Moira Burke, Cameron Marlow, and Thomas Lento. 2010. Social network activity and social well-being. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1909–1912.

[14] CARICOM. 2020. Single ICT Space, cyber security for discussion at ICT officials' meeting. https://caricom.org/single-ict-space-cyber-security-for-discussion-at-ict-officials-meeting/

[15] Hongliang Chen, Christopher E Beaudoin, and Traci Hong. 2016. Protecting oneself online: The effects of negative privacy experiences on privacy protective behaviors. *Journalism & Mass Communication Quarterly* 93, 2 (2016), 409–429.

[16] Hongliang Chen, Christopher E Beaudoin, and Traci Hong. 2017. Securing online privacy: An empirical test on Internet scam victimization, online privacy concerns, and privacy protection behaviors. *Computers in Human Behavior* 70 (2017), 291–302.

[17] Hichang Cho, Jae-Shin Lee, and Siyoung Chung. 2010. Optimistic bias about online privacy risks: Testing the moderating effects of perceived controllability and prior experience. *Computers in Human Behavior* 26, 5 (2010), 987–995.

[18] Alina Doodnath. 2018. High Court rules to decriminalise sex between consenting adult males | Loop Trinidad & Tobago. https://tt.loopnews.com/content/high-court-rules-decriminalise-consenting-sex-between-adult-males

[19] Martin Fishbein and Icek Ajzen. 2011. *Predicting and changing behavior: The reasoned action approach*. Psychology press.

[20] Donna L Floyd, Steven Prentice-Dunn, and Ronald W Rogers. 2000. A meta-analysis of research on protection motivation theory. *Journal of applied social psychology* 30, 2 (2000), 407–429.

[21] Jamaica Gleaner. 2017. Man charged for 'revenge porn', accused of posting ex-girlfriend's nude pics on social media. https://jamaica-gleaner.com/article/news/20170916/man-charged-revenge-porn-accused-posting-ex-girlfriends-nude-pics-social-media

[22] Li-tze Hu and Peter M Bentler. 1999. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural equation modeling: a multidisciplinary journal* 6, 1 (1999), 1–55.

[23] James Jaccard and Choi K Wan. 1995. Measurement error in the analysis of interaction effects between continuous predictors using multiple regression: Multiple indicator and structural equation approaches. *Psychological bulletin* 117, 2 (1995), 348.

[24] Jurjen Jansen and Paul Van Schaik. 2017. Comparing three models to explain precautionary online behavioural intentions. *Information & Computer Security* (2017).

[25] Jialun Aaron Jiang, Morgan Klaus Scheuerman, Casey Fiesler, and Jed R Brubaker. 2021. Understanding international perceptions of the severity of harmful content online. *PloS one* 16, 8 (2021), e0256762.

[26] Allen C Johnston and Merrill Warkentin. 2010. Fear appeals and information security behaviors: An empirical study. *MIS quarterly* (2010), 549–566.

[27] Dan J Kim, Charles Steinfield, and Ying-Ju Lai. 2008. Revisiting the role of web assurance seals in business-to-consumer electronic commerce. *Decision Support Systems* 44, 4 (2008), 1000–1015.

[28] Hanna Krasnova and Natasha F Veltri. 2010. Privacy calculus on social networking sites: Explorative evidence from Germany and USA. In *2010 43rd Hawaii international conference on system sciences*. IEEE, 1–10.

[29] Robert LaRose, Ying Ju Lai, Ryan Lange, Bradley Love, and Yuehua Wu. 2005. Sharing or piracy? An exploration of downloading behavior. *Journal of Computer-Mediated Communication* 11, 1 (2005), 1–21.

[30] Yao Li, Alfred Kobsa, Bart P Knijnenburg, M-H Carolyn Nguyen, et al. 2017. Cross-Cultural Privacy Prediction. *Proc. Priv. Enhancing Technol.* 2017, 2 (2017), 113–132.

[31] Han Lin, William Tov, and Lin Qiu. 2014. Emotional disclosure on social networking sites: The role of network structure and psychological needs. *Computers in Human Behavior* 41 (2014), 342–350.

[32] Sebastian Linxen, Christian Sturm, Florian Brühlmann, Vincent Cassau, Klaus Opwis, and Katharina Reinecke. 2021. How WEIRD is CHI?. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–14.

[33] Wainer Lusoli, Margherita Bacigalupo, Francisco Lupiáñez-Villanueva, Norberto Nuno Gomes de Andrade, Shara Monteleone, and Ioannis Maghiros. 2012. Pan-European survey of practices, attitudes and policy preferences as regards personal identity data management. *JRC Scientific and Policy Reports, EUR* 25295 (2012).

[34] May O Lwin, Benjamin Li, and Rebecca P Ang. 2012. Stop bugging me: An examination of adolescents' protection behavior against online harassment. *Journal of adolescence* 35, 1 (2012), 31–41.

[35] Mark Lyndersay. 2021. Considering Caribbean data protection progress. https://technewstt.com/bd1308-caribbean-data-protection/

[36] James E Maddux and Ronald W Rogers. 1983. Protection motivation and self-efficacy: A revised theory of fear appeals and attitude change. *Journal of experimental social psychology* 19, 5 (1983), 469–479.

[37] Naresh K Malhotra, Sung S Kim, and James Agarwal. 2004. Internet users' information privacy concerns (IUIPC): The construct, the scale, and a causal model. *Information systems research* 15, 4 (2004), 336–355.

[38] D Harrison McKnight, Vivek Choudhury, and Charles Kacmar. 2002. Developing and validating trust measures for e-commerce: An integrative typology. *Information systems research* 13, 3 (2002), 334–359.

[39] Norshidah Mohamed and Ili Hawa Ahmad. 2012. Information privacy concerns, antecedents and privacy measure use in social networking sites: Evidence from Malaysia. *Computers in Human Behavior* 28, 6 (2012), 2366–2375.

[40] Loop News. 2018. Guyanese cop charged for sharing ex-girlfriend's nudes online | Loop Trinidad & Tobago. *Loop News* (2018). https://tt.loopnews.com/content/guyanese-man-charged-sharing-ex-girlfriends-nudes-online

[41] Fayika Farhat Nova, Michael Ann DeVito, Pratyasha Saha, Kazi Shohanur Rashid, Shashwata Roy Turzo, Sadia Afrin, and Shion Guha. 2021. " Facebook Promotes More Harassment" Social Media Ecosystem, Skill and Marginalized Hijra Identity in Bangladesh. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–35.

[42] Cybercrime Programme Office of the Council of Europe. 2019. Report on the Regional Conference on Cybercrime Strategies and Policies and features of the Budapest Convention for the Caribbean Community. https://rm.coe.int/3148-1-1-3-final-report-dr-reg-conference-cy-policies-caribbean-comm-1/168098fb6c
[43] UN Office of the High Commissioner. 2017. UN experts urge States and companies to address online gender-based abuse but warn against censorship. https://www.ohchr.org/EN/NewsEvents/Pages/DisplayNews.aspx?NewsID=21317
[44] Jessica A Pater, Moon K Kim, Elizabeth D Mynatt, and Casey Fiesler. 2016. Characterizations of online harassment: Comparing policies across social media platforms. In *Proceedings of the 19th international conference on supporting group work*. 369–374.
[45] Andrew Perrin and Monica Anderson. [n. d.]. Share of U.S. adults using social media, including Facebook, is mostly unchanged since 2018. https://www.pewresearch.org/fact-tank/2019/04/10/share-of-u-s-adults-using-social-media-including-facebook-is-mostly-unchanged-since-2018/
[46] Sandra Petronio. 2010. Communication privacy management theory: What do we know about family privacy regulation? *Journal of family theory & review* 2, 3 (2010), 175–196.
[47] Elissa M Redmiles, Jessica Bodford, and Lindsay Blackwell. 2019. "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*, Vol. 13. 405–416.
[48] Ronald W Rogers and Donald L Thistlethwaite. 1970. Effects of fear arousal and reassurance on attitude change. *Journal of personality and social psychology* 15, 3 (1970), 227.
[49] Morgan Klaus Scheuerman, Stacy M Branham, and Foad Hamidi. 2018. 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 (2018), 1–27.
[50] Morgan Klaus Scheuerman, Jialun Aaron Jiang, Casey Fiesler, and Jed R Brubaker. 2021. A Framework of Severity for Harmful Content Online. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*.
[51] Sarita Schoenebeck and Lindsay Blackwell. 2021. Reimagining Social Media Governance: Harm, Accountability, and Repair. *SSRN (July 29, 2021)* (2021).
[52] Sarita Schoenebeck, Oliver L Haimson, and Lisa Nakamura. 2020. Drawing from justice theories to support targets of online harassment. *new media & society* (2020), 1461444820913122.
[53] Ruth Shillair, Shelia R Cotten, Hsin-Yi Sandy Tsai, Saleem Alhabash, Robert LaRose, and Nora J Rifon. 2015. Online safety begins with you and me: Convincing Internet users to protect themselves. *Computers in Human Behavior* 48 (2015), 199–207.
[54] Troy Smith and Nikolaos Stamatakis. 2021. Cyber-victimization Trends in Trinidad & Tobago: The Results of An Empirical Research. *International Journal of Cybersecurity Intelligence & Cybercrime* 4, 1 (2021), 46–63.
[55] Ashkan Soltani. 2019. Abusability Testing: Considering the Ways Your Technology Might Be Used for Harm. In *Enigma 2019 (Enigma 2019)*. USENIX Association, Burlingame, CA. https://www.usenix.org/node/226468
[56] Angelika Strohmayer, Julia Slupska, Rosanna Bellini, Lynne Coventry, Tara Hairston, and Adam Dodge. 2021. Trust and Abusability Toolkit: Centering Safety in Human-Data Interactions. (2021).
[57] Christian Sturm, Alice Oh, Sebastian Linxen, Jose Abdelnour Nocera, Susan Dray, and Katharina Reinecke. 2015. 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*. 2425–2428.
[58] Dhanaraj Thakur. 2018. How do ICTs mediate gender-based violence in Jamaica? *Gender & Development* 26, 2 (2018), 267–282.
[59] Sabine Trepte, Leonard Reinecke, Nicole B Ellison, Oliver Quiring, Mike Z Yao, and Marc Ziegele. 2017. A cross-cultural perspective on the privacy calculus. *Social Media+ Society* 3, 1 (2017), 2056305116688035.
[60] Blase Ur and Yang Wang. 2013. A cross-cultural framework for protecting user privacy in online social media. In *Proceedings of the 22nd International Conference on World Wide Web*. 755–762.
[61] José Van Dijck. 2012. Facebook as a tool for producing sociality and connectivity. *Television & new media* 13, 2 (2012), 160–176.
[62] Viswanath Venkatesh, Susan A Brown, Likoebe M Maruping, and Hillol Bala. 2008. Predicting different conceptualizations of system use: The competing roles of behavioral intention, facilitating conditions, and behavioral expectation. *MIS quarterly* (2008), 483–502.
[63] Luke Vincett. 2019. The Intersection of Race and Sexuality: An Interview with Jason Jones | Chambers Diversity. https://diversity.chambers.com/articles-media/the-intersection-of-race-and-sexuality-and-the-fight-for-lgbt-rights-in-the-caribbean-an-interview-with-jason-jones/
[64] Emily A Vogels. 2021. The state of online harassment. *Pew Research Center* 13 (2021).
[65] Pamela Wisniewski, Haiyan Jia, Na Wang, Saijing Zheng, Heng Xu, Mary Beth Rosson, and John M Carroll. 2015. 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*. 4029–4038.
[66] Irene Woon, Gek-Woo Tan, and R Low. 2005. A protection motivation theory approach to home wireless security. (2005).
[67] Liang Wu, Fred Morstatter, Kathleen M Carley, and Huan Liu. 2019. Misinformation in social media: definition, manipulation, and detection. *ACM SIGKDD Explorations Newsletter* 21, 2 (2019), 80–90.
[68] Heng Xu, Sumeet Gupta, Mary Beth Rosson, and John M Carroll. 2012. Measuring mobile users' concerns for information privacy. (2012).
[69] Seounmi Youn. 2005. Teenagers' perceptions of online privacy and coping behaviors: a risk–benefit appraisal approach. *Journal of Broadcasting & Electronic Media* 49, 1 (2005), 86–110.

A APPENDIX

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
- Social media companies would tell the truth and fulfill promises related to the information provided by me
- I 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 5: 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	Identity Theft	Your identity being at risk of theft online	0.700
AVE: 0.721	Fraud	Being a victim of fraud	0.695
α: 0.83	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	Identity Theft	Your identity being at risk of theft online	0.802
AVE: 0.794	Fraud	Being a victim of fraud	0.757
α: 0.87	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	Identity Theft	Your identity being at risk of theft online	0.896
AVE: 0.898	Fraud	Being a victim of fraud	0.900
α: 0.94	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	2FA	Use Security controls (such as two factor authentication)	0.834
AVE: 0.872	Security Checkup	Complete a Security checkup	0.912
α: 0.92	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	2FA	Use Security controls (such as two factor authentication)	0.917
AVE: 0.880	Security Checkup	Complete a Security checkup	0.916
α: 0.94	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	Trust1	Social media companies would be trustworthy in handling my information	0.828
AVE: 0.815	Trust2	Social media companies would tell the truth and fulfill promises related to the information provided by me	0.837
α: 0.91	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 usage of my information	0.753
	Trust5	Social media companies are always honest with customers when it comes to using the information that I would provide	0.849

Table 6: 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	Rumors	A person spreading malicious rumors about you on social media	0.773
AVE: 0.754	Explicit Photos	A person taking sexual photos of you without your permission and sharing them on social media	
α: 0.81	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	Rumors	A person spreading malicious rumors about you on social media	0.811
AVE: 0.817	Explicit Photos	A person taking sexual photos of you without your permission and sharing them on social media	
α: 0.88	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	Rumors	A person spreading malicious rumors about you on social media	0.882
AVE: 0.856	Explicit Photos	A person taking sexual photos of you without your permission and sharing them on social media	0.804
α: 0.93	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	Reporting - on platform	Report harassment on the platform	0.894
AVE: 0.899	Reporting - to authorities	Report harassment to the authorities (e.g. the police or build a case with a lawyer)	0.889
α: 0.96	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	Reporting - on platform	Report harassment on the platform	0.897
AVE: 0.871	Reporting - to authorities	Report harassment to the authorities (e.g. the police or build a case with a lawyer)	0.851
α: 0.944	Privacy Policy	Seek legal protection from the platform (e.g. privacy policy)	0.825
	Hide Comment	Hide potentially offensive comments/content	0.903
	Report Fake Profile	Report potentially fake profile (I.e online impersonation)	0.892
	Delete Comment	Delete offensive comments	0.854

Table 7: 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	3rd Parties	Your information was shared with third parties without your agreement	0.710
AVE: 0.758	Ads	Your information was used to send you unwanted commercial offers/ads	0.768
α: 0.83	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	3rd Parties	Your information was shared with third parties without your agreement	0.911
AVE: 0.836	Ads	Your information was used to send you unwanted commercial offers/ads	0.870
α: 0.87	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	3rd Parties	Your information was shared with third parties without your agreement	0.876
AVE: 0.854	Ads	Your information was used to send you unwanted commercial offers/ads	0.835
α: 0.91	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	Delete Post	Delete a post	0.785
AVE: 0.887	Hide Problematic Content	Hide or restrict content from particular friend/connection	0.875
α: 0.94	Unfriend	Unfriend/ Remove Connections	0.920
	Block Friend	Block/Remove Followers	0.921
	Reject Friend Request	Reject friends/ Delete Requests	0.924
Behavioral Intention	Delete Post	Delete a post	0.831
AVE: 0.896	Hide Problematic Content	Hide or restrict content from particular friend/connection	0.859
α: 0.95	Unfriend	Unfriend/ Remove Connections	0.933
	Block Friend	Block/Remove Followers	0.912
	Reject Friends	Reject friends/ Delete Requests	0.941

Table 8: 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
AVE: 0.759	Reputation	Your reputation being damaged	0.831
α: 0.87	Relationships	Your relationships with friends or family being damaged	0.819
	Physical	Your personal safety being at risk	0.822
	In-Person Stalking	Someone using your information to stalk you in person	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
AVE: 0.810	Reputation	Your reputation being damaged	0.886
α: 0.90	Relationships	Your relationships with friends or family being damaged	0.847
	Physical	Your personal safety being at risk	0.828
	In-Person Stalking	Someone using your information to stalk you in person	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
AVE: 0.887	Reputation	Your reputation being damaged	0.935
α: 0.95	Relationships	Your relationships with friends or family being damaged	0.867
	Physical	Your personal safety being at risk	0.909
	In-Person Stalking	Someone using your information to stalk you in person	0.880
Response Efficacy	Support	Seek Support (communal/offline e.g. talking to a friend)	0.893
AVE: 0.856	Advice	Ask somebody (e.g., friends, family) what I should do	0.827
α: 0.89	Safety Check	Perform safety check online	0.846
Behavioral Intention	Support	Seek Support (communal/offline e.g. talking to a friend)	0.939
AVE: 0.873	Advice	Ask somebody (e.g., friends, family) what I should do	0.898
α: 0.89	Safety Check	Perform safety check online	0.789