COGNITION, BEHAVIOR AND CYBERSECURITY

EDITED BY: Paul Watters, Dr Nalin Asanka Gamagedara Arachchilage,
David Maimon and Richard Keith Wortley

PUBLISHED IN: Frontiers in Psychology

frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88971-412-4
DOI 10.3389/978-2-88971-412-4

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: frontiersin.org/about/contact

COGNITION, BEHAVIOR AND CYBERSECURITY

Topic Editors:
Paul Watters, Macquarie University, Australia
Dr Nalin Asanka Gamagedara Arachchilage, La Trobe University, Australia
David Maimon, Georgia State University, United States
Richard Keith Wortley, University College London, United Kingdom

Citation: Watters, P., Arachchilage, N. A. G., Maimon, D., Wortley, R. K., eds. (2021). Cognition, Behavior and Cybersecurity. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-88971-412-4

Table of Contents

frontiers in Psychology

EDITORIAL
published: 05 August 2021
doi: 10.3389/fpsyg.2021.728132

Editorial: Cognition, Behavior and Cybersecurity

Paul Watters[1]*, *Nalin Asanka Gamagedara Arachchilage*[2], *David Maimon*[3] *and Richard Keith Wortley*[4]

[1] Department of Security Studies and Criminology, Macquarie University, Sydney, NSW, Australia, [2] Department of Computer Science and Information Technology, La Trobe University, Melbourne, VIC, Australia, [3] Department of Criminal Justice and Criminology, Georgia State University, Atlanta, GA, United States, [4] Department of Security and Crime Science, University College London, London, United Kingdom

Keywords: cybersecurity, cognition, behaviour, malware, phishing

Editorial on the Research Topic

Cognition, Behavior and Cybersecurity

Cybersecurity appears to be the ultimate paradox: while cybersecurity budgets are increased every year, and a vast array of new security products and services appear in the market, cyber attacks have been increasing in scale and scope every year. 2020 will perhaps be remembered as the "Year of Ransomware" as malware authors rendered useless every technical attempt to block them from attacking critical systems and data.

In this Research Topic, we have tried to present an alternative but highly complementary view to the almost total focus on purely technical solutions in cybersecurity, namely—that cybersecurity attacks ultimately succeed because they target the cognitive and behavioural vulnerabilities of ordinary users, and that for attacks to be prevented (at best) or mitigated (at least), user-focused techniques must be researched, fostered, and developed.

The small but growing band of dedicated researchers and practitioners in human factors in cybersecurity is making real inroads into developing a holistic view on how fundamental psychological principles—cognition, behaviour, perception, motivation, and emotion, to name but a few—can be readily understood within a sociotechnical context to be the primary basis for embracing a security-by-design philosophy.

Humans are complex beasts. They are motivated by a range of conscious factors and unconscious biases to make decisions that are highly exploitable by cybercriminals. Phishing texts, for example, are carefully designed to create a sense of urgency in the receiver, while malware delivery relies on the routinised habit of clicking on links. More generally, scammers exploit our inability to reconcile conflicting information in time-pressured circumstances, and our susceptibility to buy overpriced commodities during a market bubble as described in greater fool theory.

If there is one conclusion that we can draw from the body of work presented in this Research Topic, it is that computer scientists, psychologists, designers, and policy makers need to work much more closely together, to create the policy settings, technical solutions, and user validation for the secure apps and trustworthy infrastructure of tomorrow. On the one hand, a very narrow and perhaps technologists' view of user behaviour lacks sophistication, and designs ignoring psychological views are prone to exploitation.

On the other hand, more behaviourally-focused cybersecurity controls (such as auditing) can lead to abstractions (such as checklisting) that often lack the empirical connection to a deep

OPEN ACCESS

Edited and reviewed by:
Stefan Sütterlin,
Hochschule
Albstadt-Sigmaringen, Germany

****Correspondence:***
Paul Watters
paul.watters@mq.edu.au

Specialty section:
This article was submitted to
Cognition,
a section of the journal
Frontiers in Psychology

Received: *20 June 2021*
Accepted: *25 June 2021*
Published: *05 August 2021*

Citation:
Watters P, Arachchilage NAG,
Maimon D and Wortley RK (2021)
Editorial: Cognition, Behavior and
Cybersecurity.
Front. Psychol. 12:728132.
doi: 10.3389/fpsyg.2021.728132

understanding of how technologies actually work. The policy settings within which systems are allowed to be developed and operated need serious attention: Europe's General Data Protection Regulation (GDPR) speaks of "Privacy By Design," and Australia's Privacy Act (1988) relies on organisations taking "reasonable steps" to protect personal data, but there are few concrete pathways or examples of how this may be achieved using psychologically valid principles. Further integration, engagement, and mutual understanding is necessary to improve system design, and ultimately, better social and commercial outcomes.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Psychology

ORIGINAL RESEARCH
published: 30 April 2020
doi: 10.3389/fpsyg.2020.00623

Technological Change in the Retirement Transition and the Implications for Cybersecurity Vulnerability in Older Adults

Benjamin A. Morrison, Lynne Coventry and Pam Briggs*

Psychology and Communication Technology Lab, Department of Psychology, Faculty of Health and Life Sciences, Northumbria University, Newcastle upon Tyne, United Kingdom

Retirement is a major life transition, which leads to substantial changes across almost all aspects of day-to-day life. Although this transition has previously been seen as *the* normative marker for entry into older adulthood, its influence on later life has remained relatively unstudied in terms of technology use and cybersecurity behaviours. This is problematic as older adults are at particular risk of becoming victims of cyber-crime. This study aimed to investigate which factors associated with the retirement transition were likely to increase vulnerability to cyber-attack in a sample of 12 United Kingdom based older adults, all of whom had retired within the past 5 years. Semi-structured, one to one interviews were conducted and subsequently analysed using thematic analysis. Six themes were identified referring to areas of loss in: social interaction, finances, day-to-day routine, feelings of competence, sense of purpose, and technology support structures. We discuss the implications of these losses for building cyber-resilience in retirees, with suggestions for future research.

Keywords: retirement transition, cybersecurity, older adults, ageing, HCI

OPEN ACCESS

Edited by:
Paul Watters,
La Trobe University, Australia

Reviewed by:
Mohammad Jabed Morshed Chowdhury,
La Trobe University, Australia
Paul van Schaik,
Teesside University, United Kingdom

****Correspondence:***
Benjamin A. Morrison
ben.morrison@northumbria.ac.uk

Specialty section:
This article was submitted to Cognition,
a section of the journal Frontiers in Psychology

Received: *26 November 2019*
Accepted: *16 March 2020*
Published: *30 April 2020*

Citation:
Morrison BA, Coventry L and Briggs P (2020) Technological Change in the Retirement Transition and the Implications for Cybersecurity Vulnerability in Older Adults. Front. Psychol. 11:623. doi: 10.3389/fpsyg.2020.00623

INTRODUCTION

Retirement is a major life transition in which nearly all aspects of life change (Salovaara et al., 2010) and has previously been seen as "*the* psychosocial marker for entry into old age" (Kloep and Hendry, 2006). The retirement transition offers both challenges and opportunities and can be seen as a period of loss, reconstruction, and renegotiation of varying aspects of life (Price, 2003; Salovaara et al., 2010; Mao et al., 2017). Some aspects of retirement, such as changes to one's socio-economic environment and choosing how to spend newly acquired free time, have been consistent for generations of existing retirees. However, the rapid development and growth of technology provides a range of novel challenges and opportunities for those currently transitioning into retirement, and for those who will retire in the future. Technology may provide benefits to retiring adults, offering a solution to difficulties in navigating the transition to retirement. Conversely, technology may lead to additional challenges for those transitioning into retirement, such as an increased vulnerability to online victimisation.

The "Baby Boomer" generation – those born between 1946 and 1964 (Young and Tinker, 2017) – are currently making the transition to retirement and are likely to be the first generation to experience the costs and benefits of this technological change. This generation has lived through a digital revolution and are likely the first retirees to have used technology for a large part of their working lives (Durrant et al., 2017). Their engagement with technology makes them the first generation who are likely to use technology before, during and well into retirement. Technology use by this generation has steadily increased over time. For example, around 50% of this age group in the United Kingdom own a smartphone and those who say they never use the internet has dropped from 49 to 29% in the last 5 years (Ofcom, 2018).

Early research into technology use in older adults (for example: Gregor et al., 2002) was based on the premise that those in retirement were not active technology users. The concept and associated language, metaphors, and behaviours were unfamiliar to them, and they did not necessarily perceive the benefits of technology use. Today's older adult population demonstrate a normative shift towards technology, with older adults eager to adopt new technology (Mitzner et al., 2010; Vaportzis et al., 2017), recognising its utility for maintaining independence for longer into older age (Lindley et al., 2008; Seifert and Schelling, 2018). Many older adults now engage with online technologies to counteract loneliness and isolation (Chopik, 2016), remain socially connected (Hutto and Bell, 2014), interact with family, and enjoy a healthier retirement (Khvorostianov et al., 2012; Juárez et al., 2018). However, these benefits are associated with known costs in terms of an increased risk of online victimisation for this population (Chakraborty et al., 2013; Sarno et al., 2017).

All digital technology users are potential victims of cyber-attacks and may even unknowingly participate in those attacks (Von Solms and Van Niekerk, 2013). Unsurprisingly, a growing body of cybersecurity research focusses on the role of the user. Here the drive is to understand the role of end-user behaviours and attitudes which range from authentication behaviours (Nicholson et al., 2013a,b), web browsing (Kisekka et al., 2015), decision making (Jeske et al., 2016), through the risky behaviours such as password sharing (Whitty et al., 2015).

Despite this growing literature base, there remains a paucity of cybersecurity research that explores the older adult user, even though this population may be at increased risk from cybersecurity threats (Grimes et al., 2007, 2010; Age-UK, 2015a; Age-Uk, 2018).

Retired, older adults are likely to be more susceptible to specific online threats. In a small scale qualitative study of three older adults, Olivier et al. (2015) suggested that they may be at particular risk of mass marketing fraud due to their pscho-social backgrounds and pre-disposing factors such as pschological vulnerability, something which was supported in a recent review of mass marketing fraud by Shao et al. (2019). Other research has identified an increased vulnerability to: telemarketing fraud (Alves and Wilson, 2008), phishing (Cho et al., 2016; Sarno et al., 2017), pension Scams (Martin and Rice, 2013; Nicholson et al., 2019), and other such targeted attacks. Those born before 1954 are also more likely to perform fewer protection behaviours and are less confident in their own ability (Jiang et al., 2016; Nicholson et al., 2019). Furthermore, they have less trust in protective resources and are more likely to be rely on others' assistance when compared to younger generations (Jiang et al., 2016). In addition, Forget et al. (2016) interviewed 15 participants, most of whom fell within the baby boomer age range, demonstrating that security problems often arise when there are disconnects between what users see as their computer security role, and what is expected of them by others. Investigating these potential sources of vulnerability to cyber-attack is important, as these attacks can have a range of negative consequences at individual, community, and national levels (Saini et al., 2012; Canetti et al., 2017). Additionally there are ethical implications for society as a whole, since the protection of vulnerable groups could be seen as a societal responsibility (Von Solms and Van Niekerk, 2013).

One challenge in this research space is the tendency to focus on chronological age as a defining characteristic of an individual. Researchers tend to classify users into arbitrary age groups such as young children (Guan and Huck, 2012), teenagers (Wittes et al., 2016; Rahman et al., 2017), late midlife (Salovaara et al., 2010), and older adults (Chakraborty et al., 2013; Hill et al., 2015), where chronological age determines group selection, sometimes to the detriment of other socio-economic or psychological variables. Yet age is not a reliable marker for any particular user attribute, thus observing individuals based on chronological age not only risks research ageism (Vines et al., 2015), but also risks underestimating the effect of substantive life events (Shultz and Wang, 2011) and the impact they may have on cybersecurity vulnerability. This is particularly true when we consider retirement. Using semi-structured interviews with eight purposefully selected, recently retired individuals, Pettican and Prior (2011) found that retirement was identified as a new life stage, with retirement being a time of significant re-adjustment, with changes influencing perceptions of both of health and wellbeing. The varying retirement trajectories that retiring individuals face are likely to be diverse and based on a range of factors such as preparedness, planning, and the socio-economic circumstances surrounding the retirement transition. Inevitably, the post-retirement experiences of older adults are likely to vary greatly.

A large scale systematic review by Barbosa et al. (2016) outlined 26 key areas associated with adjustment into retirement, with a range of positive and negative outcomes identified in an emerging literature base. Their review demonstrates that the majority of research regarding the retirement transition revolves around physical and psychological health; however, a number of other factors are important when considering the transition from the workplace into retirement. Existing cybersecurity literature has so far failed to address the impact of retirement as a major life transition on technology use and cybersecurity vulnerability. However, some of the factors of retirement adjustment, as seen in Barbosa et al. (2016), might logically be related to cybersecurity vulnerabilities. For example, one factor identified within their review relates to finances and how financial strength influences retirement outcomes. Although it is likely that being in a strong financial position is beneficial in general when transitioning into retirement, it is likely that such financial strength might lead to increased targeting by cyber-criminals (Oliveira et al., 2017).

Another such example relates to social integration, within the Barbosa et al. (2016) review, this factor is seen to have a mixture of findings in terms of whether it is positive or neutral in its relationship to retirement adjustment. In an increasingly technological world, those who are isolated in older age may turn to social media and technology to reduce loneliness and isolation (Nowland et al., 2018), something which may open the door to certain cyber-attacks such as romance scams (Buchanan and Whitty, 2014; Whitty, 2017) or social media based cyber-attacks (Jang-Jaccard and Nepal, 2014). It is clear that a number of the factors associated with retirement might influence technology usage and as such may contribute towards cybersecurity vulnerability in retirement; however, there is presently very little research in this area. In a 4-week qualitative study, Durrant et al. (2017) investigated technology use in six recently retired older adults. They identified that older adult's adoption and use of internet-enabled devices had a significant presence in aiding with the transition into retirement. Although they identified some security concerns relating to this technology use, very little research has looked at how these transitions, and the changing nature of this technology use might influence vulnerability to cyber-attacks.

Henkens et al. (2017) outline three areas in need of research, linking technology use to the retirement transition: (1) the impact of technology on financial preparedness, (2) the impact of technology on facilitating a better work-life balance and thus allowing people to work longer, and (3) the way technological advancements lead to improvements in psychological adjustment. While these research areas are undoubtedly important to aiding and understanding the retirement transition, we suggest that a fourth area is important but remains unaddressed – how the retirement transition can lead to increased cybersecurity vulnerability.

In this study, we aim to explore how interaction with technology changes, in both online and offline environments, as a result of the retirement transition. Furthermore, we draw from a cybersecurity literature to demonstrate how these changes are associated with the implicit cybersecurity vulnerabilities that we see in older adult populations. In doing so, we seek to provide a foundation for future work that investigates how retirement, as a major life transition, might act as an antecedent to cybersecurity vulnerabilities in post-retirement life.

MATERIALS AND METHODS

Participants

Face to face and online sampling methods were used to search for eligible participants. A post was placed on Facebook in January 2018, which asked directly for participants, but also requested that people snowball on the recruitment information to anyone who might be eligible to take part. Once potential participants had contacted the research team, an interview was arranged at the participant's home. A total of 12 participants from the North East England, United Kingdom, took part in the study. Although we did not originally specify a specific number of participants prior to conducting the study, due to the difficulty in establishing such measures (Levitt et al., 2018), we ceased data collection at the point which we reached data saturation, "the point at which no new themes or information arose" (Guest et al., 2006). The number of participants involved in this study is in-line with a range of existing qualitative studies in the area of cybersecurity (Olivier et al., 2015; Durrant et al., 2017; Fujs et al., 2019). The sample consisted of seven females (aged 59–74 years) and five males (aged 53–68 years) (see **Table 1**) who met the criteria; that they had experience in using technology, and had retired within the past 5 years. These participants were from a diverse range of backgrounds with varying technical expertise ranging from careers in retail through to engineering.

Materials

An interview schedule was created based on factors identified within the Barbosa et al. (2016) systematic review of retirement adjustment. First, the 26 factors of retirement research were screened to identify which factors were likely to influence changes in technology usage or increase cybersecurity vulnerability in any way. Following this review, six major factors were identified: (i) social situation, (ii) online/technology adoption, (iii) identity transitions, (iv) psychological wellbeing/personality change, (v) support structures, and (vi) financial change. Of the 26 factors attributed to retirement adjustment, these six factors were seen

TABLE 1 | A descriptive overview of participants.

Participants	Age (years)	Retired length	Sex	Pre-retirement work
P1	59	4 months	Female	Worked as a nurse for most of her career before moving into a role as a manager. She is married and living with a chronic health condition.
P2	60	2.5 years	Male	Project manager who worked for BT is married to P6.
P3	68	3 years	Male	Worked as an engineer for 30 years before spending 15 years in project management at BT.
P4	60	4.5 years	Male	Worked in a ministerial role relating to schools' funding.
P5	59	6 months	Female	Worked as a technical support operator giving IT support and building IT systems.
P6	67	2.5 years	Female	Retail shop assistant. Is married to P2.
P7	66	5 years	Female	Worked as team leader in a café on a University campus.
P8	63	2 years	Female	Worked as a nurse for most of her career but ended career being a manager to other nurses. Is married to P9.
P9	65	1 month	Male	Worked as a security engineer fitting and maintaining ATMs, is married to P8.
P10	62	18 months	Fe	Retired 4 years ago but following a 3-month break this participant returned to work as a FE school vice-principal. Retired again 18 months ago.
P11	74	3 years	Female	Worked as a GP practice manager.
P12	53	2 years	Male	Worked as a self-employed charity worker.

not only to be related to retirement, but likely to have some influence on technology usage. Interviews began by asking the participant to outline what had happened to them during their transition from the workplace into retirement. The interview then went on to ask what the biggest changes were across their retirement transition and what impact these changes had had on their day-to-day lives. The known factors, which were likely to change as a result of the retirement transition, were included as prompts within the interview to stimulate discussion. Prompts around computer and technology use were also included to stimulate relevant discourse.

Procedure

Ethical approval was obtained from the psychology ethics board within the University of Northumbria at Newcastle upon Tyne. Semi-structured interviews were conducted in participants' own homes, so that their devices were present and could be used to stimulate discussion. During the interview, if a participant had mentioned a specific digital device, or if it was present in the room, they were asked if they could talk about how they used the device, and if they would show the researcher the sorts of apps or software packages that they typically used. Interviews took approximately 1 h and were structured around three main topics: (1) what participants experienced in the lead up to their retirement, (2) what the participant saw as the biggest changes to their life over this period, and the reasons behind this, and (3) if and how the participants online behaviour had changed across the retirement transition.

FINDINGS AND DISCUSSION

Analysis Procedure

The data was analysed using NVivo 11 software by the first author using a "soft template" (King, 1998), where the six *a priori* themes were used as an initial coding guide in accordance with an iterative template analysis (Brooks et al., 2015). Halfway through coding, and following discussion of the data with other authors, these themes were revised, resulting in a final set of six themes described below.

Themes

During the coding phases, it became clear that emerging themes generally related to areas of *loss* rather than change, and thus the initial template was revised to reflect this. Losses were typically accompanied by compensatory behaviours, which, in some cases, contributed towards cybersecurity vulnerabilities in older adults. Each area of loss also carried with it emotional implications, which may have been the driving force behind attempts to remedy these losses. Each of these "losses" is discussed below, followed by an explicit articulation of the cybersecurity implications of that loss.

Changes in Social Interaction

Upon leaving the workplace, an individual's social infrastructure changes and in most cases, the social and emotional support from workplace colleagues is lost. Our participants had typically made the transition from working full time (around 37.5 h per week) to being fully retired and this drop-off in working hours led to rapid social change. For some, the loss of colleague interaction occurred immediately, as they had chosen not to maintain contact with colleagues. Others described their attempts to keep in touch with colleagues, although this too gradually deteriorated over a period of time.

> *P6: I did socialise with people from work. Not a lot, but once I had retired, that got less and less, and it was sort of once a month I would speak to the girls from work, then it sort of got to once every couple of months and people kept in touch with me once I retired, but then as the months went on it got less and less and now. . .well two and a half years now since I retired, I virtually don't see anybody from work at all.*

Nearly all participants described a vacuum in their social infrastructure. For many, social interaction had revolved almost entirely around work colleagues, and this meant that rebuilding social interaction post-retirement was difficult.

> *P1: A lot of nursey people do hang about with nurses, so when you stop doing that you find that trying to spread your group of friends a bit wider is a bit tricky.*

This loss led to increased feelings of isolation and loneliness in many participants. This was especially apparent when the loss of social interaction felt like a slow process of neglect.

> *P6: When you are at work there is lots going on "oh we're planning a night out on Friday, are you going to come?" Once you are out of the picture I think you are soon forgotten.*

Generally, social loss was seen as highly negative. This finding is not surprising and is entirely consistent with the observations of Kloep and Hendry (2006), who demonstrated that people who become attached to colleagues are unhappy at losing them as part of their social infrastructure. Dorfman (1992) argued that the loss of colleague interaction was rated as the most negative aspect of retirement. Nahum-Shani and Bamberger (2009) found that in those with a large number of working hours, retirement not only led to a loss of colleagues, but also led to a decrease in emotional support overall, i.e., work friends who were previously strong emotional support structures were no longer available to the retired individual. It may be that people look to refill this social loss not only for interaction, but also for the emotional support it provided.

Renewing Social Interaction

In compensation, participants sought out new social opportunities via taking on new hobbies, joining groups, volunteering, and providing support for the family.

> *P5: One of the purposes behind me deciding to do some volunteering, I picked the library specifically so that I could meet people in the Low Fell area, because I have never had children, I have never done the school gate business, I don't really know anyone, apart from immediate neighbours.*

For those who were married, a renegotiation of the marital relationship was required to establish whether this loss of colleague interaction would be replaced by more time spent together.

> *P8: Now he is retired, we are kind of like sorting out how we can get on with life as two people again, instead of one working and one not working.*

They described turning to technology to facilitate social interaction with those outside of their immediate home. Some started WhatsApp messaging; others joined online social networks as a means to meet new people.

> *P8:. . .That has definitely increased since I have retired, the texting and the emailing and the WhatsApp.*

> *P4:. . .Because the children are growing up there is a lot more sending photos, because my nephews and nieces, most of them have got kids now, so again its photos of the kids, a LOT more texting, a lot more texting actually because I have a lot more time to do it.*

Some participants described the way their social interaction now revolved around family life. They explained how their family members had bought them devices or encouraged them to use online social networks.

> *P8: She gave me this phone so that I could receive photographs and so that she could Skype me. Not my Skype her, but her Skype me. And. . .FaceTime? Is it FaceTime? So that kind of thing.*

In line with these findings; Peek et al. (2016) found that it was common for families to buy devices for their older relatives and in general those within the individuals social support structure facilitated the use of technology. However, within our sample, not everyone felt competent or confident in the use of such devices and some reported that there were emotional implications in using these devices such as general fear and anxiety around unfamiliar device usage.

Vulnerabilities Arising From the Loss of Social Interaction

As retirees seek to build a new social life, they may turn to technology and social networking platforms to build up communication with family, find new people with common interests or new ways to express themselves (Tosun, 2012). However, online social networks are recognised as one of the biggest emerging threats to cybersecurity and privacy (Jang-Jaccard and Nepal, 2014). Increasingly, these outlets are being used to spread malware, gain information for identity theft or to seed romance scams. As retirees take to online social networks, they may increase their vulnerability to attack, particularly if they are not competent or confident with the technologies they are using. Retirees must ensure that they have anti-malware up to date and active if using such sites as they become prone to phishing attacks, romance scams, and grandparent attacks (Alves and Wilson, 2008; Age-UK, 2015b).

Changes in Finances

Most participants experienced an immediate loss in income upon leaving the workplace. Some were financially prepared, meaning that their salary was substituted by a good pension, and reduced outgoings, e.g., being mortgage free. Finances varied among the participants with a few reporting that they were financially better off overall since retiring. More commonly, however, people experienced a large loss in their income, which resulted in changes to their financial behaviour and attitudes.

> *P11: One of the biggest transitions and the worst part for me is the lack of money. Um, suddenly going down from having a salary to a pension that worked out to be much less than I believed I would have received.*

Participants had to change their lifestyle in order to live within their means. Participants reported managing their spending more carefully; being careful not to overspend and ensuring they sought value for money.

> *P5: Oh god yeah, my pension is about a 1/3 of what I used to get paid and although I have a decent amount of savings, I am finding it very interesting having to actually watch what I spend on a month by month basis.*

> *P8: I am more careful with my money because I don't have the disposable income I used to have, and it's fun in a way hunting for a bargain and when you go out you get concessions because of our age. It just makes you look at money in a different way.*

Financial loss had an impact on multiple aspects of life and had consequences for other retirement related losses. For example, the need to limit expenditure further, amplified the social interaction losses as social events and club memberships were perceived as too expensive.

> *P7: I definitely don't socialise like I used to, because I was out every week, every single week, I was out every weekend. Q: Why don't you do that anymore?. . .I think of the money, do you know what I mean? Because when I worked [. . .] you would probably spend £70 on a night out and that is a lot of money when you are on your pension, that is nearly your week's shopping.*

> *P11: Until I left work I was a member of xxxx cricket club, but I can't afford that any longer, so I always went to matches as much as I could at xxxx, but I don't do that anymore, I just can't afford it.*

For those without a car, reliance on public transport (often perceived as inaccessible or inflexible) led to further forms of isolation.

> *P11: It means I can't afford to run a car any longer so that is a big change. I have had a car for many, many years um, certainly since my late 20s I've always had a car that I've run, even when I was really hard up, then it was still easier to do it than it would be now, so yeah that's one of the real big disadvantages.*

This resonates with the findings of Davey (2007), who reported a range of negative outcomes associated with the loss of a car, including difficulties in carrying out day-to-day tasks, going to see friends or shopping without assistance. These findings are also highlighted in Luiu et al. (2017); their review of the literature surrounding the implications of losing access to personal transport in older age.

Some participants were understandably frightened by loss of income.

> *P9: So that has gone down just to my two pensions. It's a bit – that's what frightens you at first and you think bloody hell, where is it, before I had the money if I wanted to go for a day out.*

Post-retirement satisfaction and happiness have both been found to relate to financial status (Choi, 2001; Kim et al., 2001; van Solinge and Henkens, 2008). Burr et al. (2011) found a strong association, in that a good, stable income led to positive affect whereas poor financial status led to negative affect.

Vulnerabilities Arising From Financial Changes

Participants who were financially comfortable, post-retirement, reported very little in the way of associated behaviour change. However, those who had experienced financial loss, reported being much more attentive towards finances, with a number of new reported behaviours, including greater interest in online banking as a means to manage finances well:

> *P4: Online banking is something I have always done but I am much tighter on, but before I retired, and I didn't really need to worry much there was always kind of enough money for what I wanted, now I have to be very careful.*

This suggests that financial loss during retirement can be protective in a cybersecurity context, as the individuals attention becomes focussed on protecting their limited resources. This is supported by existing literature which indicates that those with a higher income are less risk and loss averse (Hjorth and Fosgerau, 2009; Sheehy-skeffington and Rea, 2017). Grable (2000) also found that those with a higher income and higher education level had a greater risk-taking propensity.

However, while those with stretched finances may engage in more protective checking behaviours, financial loss could lead to other cyber-vulnerabilities, e.g., through using second-hand technology, something which is especially problematic if such behaviours are not perceived to be risky at the time. One participant (P4) reported how he experienced a large financial loss following retirement, and had bought an old used laptop for £80 as well as purchasing anti-virus software from local paid IT help. He described this as his "clunky laptop." It had an outdated operating system but was his main portal for accessing information, exchanging emails and downloading information from the Internet.

Financial loss could, thus, lead to unsafe behaviours such as purchasing of used, outdated or inherently unsafe devices. Some who are struggling financially may rely on hand-me-down devices from friends or relatives in an attempt to avoid spending precious financial resources on new technology. A lack of financial stability may also hinder people from buying security software, paying for IT help when required, and relying on those available to the individual (Dimond et al., 2010; Nicholson et al., 2019) regardless of their ability to provide good technical support.

Loss of Sense of Purpose

The workplace can provide people with a sense of purpose or strong professional identity. Participants described feelings of loss around their former role, with some saying they no longer felt that they had a place in society, while others described feelings of guilt at no longer being in useful employment.

Some participants had worked in specific job roles for their entire working lives and their working role had become a large part of their self-identity. Upon retirement, they were forced to re-assess their identity, and this could be difficult.

> *P1: It does kind of dominate your life, it sounds pathetic really, you are even a nurse when you are not at work, and you know. [. . .] It's not being a nurse anymore, I find that quite odd.*

> *P5: I have always really defined myself in a large part by what I do, and I suppose work was always very important to me because it took a lot of my life and now I don't do that anymore I am JUST retired. . .I am JUST. . .*

Conversely, retirement had relatively little identity impact for those who were unhappy in their pre-retirement roles.

> *P4: I think the difference is because I didn't like my job for the last few years, I didn't proudly identify myself with the role, it was "this is what I am doing to earn enough money" and that's how it felt. And so, I didn't. . .it wasn't like losing an element of my identity, or the element of my identity that I lost didn't like anyway.*

Role theory is a transitional theory that relates to specific roles gained and lost across the life course and may be particularly helpful when investigating the loss of a sense of purpose in recent retirees (Wang et al., 2011). Retirement acts as a role-transition (Wang, 2007) which may lead to a loss in feelings of purpose. Kim and Moen (2002) outline how, from a "role-enhancement" perspective, the loss of a career leads to feelings of "role loss" which in turn drive feelings of psychological distress and loss of morale. Alternatively, leaving a role that the individual is unhappy with, can lead to a reduction in "role strain" (Kim and Moen, 2002).

The loss of a workplace role can damage one's self-identity (Osborne, 2012) or one's self-esteem (Bleidorn and Schwaba, 2018) although this can depend upon the way the exit from the organisation is handled (Damman et al., 2015). Our participants used emotive language when discussing role loss following retirement, using terms such as "feeling useless," experiencing "crises," or likening the experience to "jumping off a cliff."

> *P5: I am having a bit of a very low-key crisis of wondering where I fit in the world, but. . .I don't think it is anything I won't get over.*

> *P6: I think the biggest change is that I felt. . .It's difficult to describe. . .not that I was useless, but I felt like I wasn't. . .that I didn't have any valuable contribution to make.*

Such distress can act as an impetus to re-fill this role loss. Participants took on a variety of new roles in retirement. If they had grandchildren living nearby, they generally reported taking on more active roles as grandparents.

> *P10:. . .The other thing that takes over when you retire, when you're a grandparent, is visiting the little ones.*

Others took on roles such as volunteering, turning to part time work or increasing the amount of time spent doing hobbies and activities. One recent retiree said he did not yet know what to do with his spare time, likening the experience to an earlier life transition, that of leaving school.

> *P9: I'm at the point now, like just before I left school not knowing what I want to do – it's like, when you leave – where are you gonna go? I'm sitting here scratching my head thinking I don't know – how long that will take I don't know.*

Thoits (2012) describes the ways that taking on a new role (e.g., volunteering) can lead to increased feelings of self-worth, renewed feelings in a sense of purpose and better physical and mental health. Volunteer roles are popular post-retirement, as they are relatively easy to obtain, are likely to involve low stress, and are typically easy to exit. However, these roles may bring new challenges and vulnerabilities.

Vulnerabilities Arising From Loss of Sense of Purpose

Participants taking on new roles were sometimes given technology responsibilities, regardless of their actual ability. As noted earlier, this is predominantly a group of "baby boomers," i.e., the first group of retirees to have had technology experience during their working lives. At times, this responsibility was accepted and at other times refused.

> *P5: Well, I have started looking after the website which is not. . .a particularly difficult job it is on a contact management system, but I do the updates on it and. . .and I look after their Facebook page as well, and I make use of my laptop a lot more than I used to.*

> *P2: The art group has asked me to manage their Facebook page for them, one of my neighbours has asked me to get involved in the Elders group in Newcastle and help them with their web development and I'm afraid I have said no to all of them, I have spent 40 years in technology and I hate it.*

Even for those without a strong knowledge of technology, new roles often led to an increase in technology use associated with communication.

> *P4: I am in constant contact with the people who run it, the chief executive if you like I am his line manager who I see once a fortnight, we exchange a LOT of texts and emails on that.*

This can be problematic for those people who are given access to systems they are ill equipped to protect. A large increase in the amount of emails that an individual handles is likely to increase an individual's exposure to email related threats such as phishing attacks. Parsons et al. (2019) suggests that those with more technology experience will outperform those with less in terms of avoiding phishing threats, but the vulnerabilities of a new "volunteer" might not be made explicit to a recruiting organisation.

A Loss of Day-to-Day Routine

Following retirement, participants found themselves without a day-to-day routine, which was sometimes associated with feelings of guilt about having so much free time.

> *P6: I think the biggest change is that you are suddenly in an environment where you aren't busy, you go to work, I went to work 5 days a week, then all of a sudden you haven't got that.*

> *P5: I find it still very hard to just not have something planned to do because I feel like I am wasting time, I feel a bit guilty.*

Osborne (2012) describes the "choice dilemmas" that can lead to feelings of angst or anxiety in retirees. Siegenthaler and Vaughan (1998) found that retired women often reported feeling guilty about engaging in recreation during retirement. Again, these changes can lead to a change in behaviour.

> *P5: I had gone from having a very structured life to suddenly having no structure and all of this spare time to do things, so I immediately set about putting structure in place, I volunteered at various things. . .*

Having more free time was a reason cited for participants taking up a range of activities: re-discovering previous hobbies, dedicating more time to existing hobbies and adopting new hobbies or activities. Again, for participants this was accompanied by an increased use of technology, as they now had more time to engage with digital devices.

> *P5: Facebook I didn't do when I worked I do a bit more of now, I watch more things on television and Chromecast, I have Netflix which I didn't have when I worked. . .I just needed more time. I didn't have time for anything like that. It really was precisely that.*

One participant outlined how boredom led to an increase in online social network participation.

> *P4: Oh yes, I didn't use it [Facebook] at all, I think it is completely new since I retired, I have more time as well, sometimes I look at Facebook because I am a bit bored.*

Tosun (2012) found that a common reason for Facebook use was to curb boredom and we found evidence of other similar online activities, driven by a need to fill the retirement hours.

> *P8: I text and WhatsApp friends as well, quite. I guess daily really, I will sort of text people and ask, how is your mum and when are we meeting up and they will WhatsApp me back and things. That has definitely increased since I have retired, the texting and the emailing and the WhatsApp. Because I have the time to do it now.*

Barnett et al. (2012) outline how retirees need to replace their working day routine with new routines in retirement for the purpose of maintaining a feeling of control and a sense of purpose over their lives. Beck et al. (2010) focussed on the initiation and maintenance of physical activity in retirement, recognising the close link between a loss of routine and a loss in a sense of purpose. Ekerdt and Koss (2016) found that routines were seen as vital by retirees for a number of reasons, one of which is to address the open-endedness of retirement and to instil a sense of purpose and meaning to one's post-retirement life.

Vulnerabilities Arising From Loss of Day-to-Day Routine

Having more free time in retirement almost inevitably meant that our participants spent more time using technology. Choi (2008) suggests that one's online routines and the way in which these routines are managed, provide opportunities for victimisation in an online environment. We noted earlier that social media use is one of the biggest emerging threats for cybersecurity (Jang-Jaccard and Nepal, 2014), but boredom can also lead to increases in things like online play which brings a number of cybersecurity concerns, particularly when that play is associated with apps downloaded onto smartphones and tablets (Lu et al., 2012; Ahvanooey et al., 2017).

Loss of Perceived Competency

Another major change that occurs following departure from the workplace is the immediate reduction in work based cognitive demands. Within this sample, participants discussed how they felt less "mentally fit" after retiring. Additionally, they reported a decline in their computer self-efficacy related to these perceptions of declining competence.

Participants described feeling cognitively "slower," and attributed these losses to their retirement transition.

> *P1: You find yourself reading the same thing over and over again and not taking anything in. I used to find that after a fortnight off [. . .] if I went back after a fortnights holidays I wouldn't be as sharp as when I went off until I had revved back up. And of course, I haven't revved up since September.*

> *P2: I'm sure that would have taken me an hour or so if I was. . .you know, before I retired. This time, I kept making mistakes and it wouldn't sort, or it wouldn't go quite how I thought it would, so I must have spent 5 h doing three sheets of A4.*

It may be that time spent in the workplace acts as cognitive protection, allowing people to "flex their mental muscles" in regard to carrying out a broad range of tasks. Evidence by Finkel et al. (2009) demonstrated that pre-retirement job roles that involve highly complex work, resulted in better cognitive functioning following the retirement transition. We know that, regardless of chronological age, adults may show more rapid cognitive decline following departure from certain workplaces and that this is linked to the complexity of the work previously undertaken (Finkel et al., 2009; Meng et al., 2017). Gordon et al. (2019) also found that older adults, regardless of chronological age, could be divided into "cognitively young" and "cognitively old" individuals and that this was reflected in their technology usage, with "cognitively older" adults using fewer apps for longer periods. These declines may not necessarily be reflective of actual cognitive decline however.

There is no doubt that actual cognitive and physical decline occurs for many and often begins before the age of 60 (Salthouse, 2009) which may in turn, may be linked to problems in mastering new technologies or even in the everyday ease-of-use of existing technologies (Hauk et al., 2018). However, people also show a number of negative self-perceptions about ageing (Sargent-Cox et al., 2012; Robertson and Kenny, 2016) which in turn can lead to doubts about competence beliefs. The perceptions of declining competence that retirees report following departure from the workplace, may instead be related to declines in self-efficacy rather than actual cognitive decline. Self-esteem gradually rises across the life course, starts to decline around the age of 50–60 years, and continues to reduce into older age (Orth and Robins, 2014). Retirement; through losses of roles, purpose and perceived competence, especially in the oldest retirees, may intensify age related declines in self-efficacy beliefs. This may be particularly problematic as declines in self-efficacy have been associated with increased cybersecurity vulnerabilities.

Vulnerabilities Arising From Loss of Perceived Competence

Seeman et al. (1999) demonstrated that in a sample of older adults, low self-efficacy beliefs led to incorrect assumptions of performance decline, i.e., older adults saw themselves as declining, even though this was not the case objectively. Torrens-Burton et al. (2017) also demonstrated a discontinuity between perceived and actual behaviour in older adults. Although information processing scores were the same, those who perceived that they had higher levels of cognitive dysfunction believed that they had performed more poorly, indicating a lower self-efficacy belief.

In a technology context, Vaportzis et al. (2017) found that older adults (aged between 65 and 76 years) had feelings of inadequacy when comparing their computer literacy with those of their peers. Similarly, Marquié et al. (2002) supported this in the context of general computer knowledge as they demonstrated that older adults (with a mean age of 68.6 years) underestimated their computer knowledge when comparing themselves to a younger sample (with a mean age of 22.6 years). They found that older adults were both less confident and felt less knowledgeable, regardless of the fact that their scores were in line with their younger counterparts. Although older adults may be capable, a perception of low self-efficacy may be damaging nonetheless. Workman et al. (2008) suggest that an individual's ability to cope with an online threat is partly based upon their self-efficacy, finding that those with lower self-esteem were more likely to engage in omissive behaviours around information security. Thus, lowered self-esteem may result in avoidance behaviours, rather than attempting to deal with threats directly.

These findings are important in the interpretation of our own data, as we found incidents where low perceived computer self-efficacy and the fear and anxiety around "doing the wrong thing" drove participants to seek sources of support, which may not always have been appropriate or safe.

> *P8:. . .My granddaughter will point me in the right direction, if I get really stuck I will say "help, I'm stuck" because I am afraid that I might do something wrong and lose everything. And I don't know how to get it all back, I am very naïve when it comes to things like that.*

Nicholson et al. (2019) has shown that older adults will often behave in just this way – showing reluctance to master new procedures and turning instead to close relatives or readily available others to fix things, without necessarily checking their credentials for undertaking the task at hand. Barnard et al. (2013) noted that having access to a particularly knowledgeable child or grandchild, may backfire, reinforcing feelings of incompetence when they see the third party confidently and competently handling technology. If older adults become reliant on others for cybersecurity support, especially if these sources are inappropriate, there is a clear risk in terms of cybersecurity vulnerability.

Loss of Technical Support Structures

Many workplaces provide technical training and support to staff members and most have appropriate policies and procedures in place. Alongside formal IT support, knowledgeable colleagues provide technical information and advice through socially constructed "shadow security" networks (Kirlappos et al., 2014). These are all lost upon retirement (Dimond et al., 2010) and our participants recognised this as an issue. They described how the workplace had provided support in the form of bulletins, updates and dedicated IT staff and described their reliance upon workplace friends and colleagues for technical support.

> *P2: Yeah. At work they are all very technical people, [. . .] so I would go to them. If I had a problem I could just phone a help desk at work. But if you phone a helpdesk when you are at home then it costs money doesn't it?*

> *P1: Yeah. I don't know who else I would ask actually [for IT help]. At work I could find anyone with an iPhone and say here this has happened, what do you think?*

Participants also described a reliance on workplace support structures to keep them updated with cybersecurity threats and to act as reminders of safe practices.

> *P1: You don't realise how much you rely on it for, there were banners going across the computer screen homepage all of the time telling you about them [threats] and to update.*

Nahum-Shani and Bamberger (2009) found that working hours were positively associated with the depth of colleague instrumental support received (support with devices) but showed how this was lost upon retirement. Instead this was replaced by advice and support from those close (non-work) friends who tended to be immediately and easily accessible – findings similar to those reported for cybersecurity advice by Nicholson et al. (2019). It appears that the options for retirees become limited, and they have to rely on what would have been their second or third choices for support.

> *P11: My youngest daughter is probably the principal person who would have helped me, but now I'm living here, and she lives in London, or just on the outskirts of London so she isn't around as much, whereas [granddaughter] lives down the road.*

Some participants reported employing paid help for IT support, as they no longer had available IT support structures at all.

> *P4: You see when you are working. . . [. . .] there are always people around to ask questions, that is one change, there aren't anymore. I suppose that is why I take the machine to him [local paid help] every now and then to get it cleaned up. . .*

People may not want to admit incompetence to family members, feeling embarrassed about their inability to deal with threats (Selwyn, 2004). It is clear, however, that the choice of support structure in retirement may result in an increase in vulnerability to cyber-attack in retirement, depending on their trustworthiness and knowledge.

Vulnerabilities Arising From Loss of Support Structures

Nicholson et al. (2019) may help to clarify the mechanisms by which a loss of support structures in older adults may lead to cybersecurity vulnerability. They posit a framework in which cybersecurity information is a result of an interplay between cyber-literacy and resource availability. For a retired individual, the legacy knowledge they acquired in the workplace is often used to guide their cybersecurity behaviour, but as this information becomes more dated, they turn to other resources for support and the acquisition of new knowledge and skills. Yet as we've seen, the post-retirement resource landscape is very variable. Some people have a wide and knowledgeable social network. Others, with more financial stability, have bought new devices and therefore have ready access to professional IT support. This pattern has been noted by Barnard et al. (2013), who notes that retirees place themselves "at risk of being left behind" which sometimes leads them to make risky decisions or rely on outdated or inappropriate advice. We see this pattern in our own data. For example, when asked about what to do if no support was immediately available, one participant explained how she might engage in behaviour outside of her comfort zone to achieve her end-goal.

> *P1: If I was confident about the website. So, if it was it was iTunes. Like the computer died [. . .] and iTunes had disappeared. I downloaded that again but with clammy hands because it had to be updated, and I am a heart in the mouth kind of IT person really.*

Surprisingly, there is relatively little in the research literature about how such challenges, and more specifically about how changes in post-retirement support structures can leave people open to attack.

OVERALL DISCUSSION

We have documented a number of losses associated with retirement and shown how these can make older adults more vulnerable to cyber-attacks. Our evidence supports the notion that retirement acts as a major life disruption and one which leads people to seek out a "new normal" (Massimi et al., 2012), i.e., a new lifestyle in which previous technological and social infrastructures are lost and are subsequently replaced with tenuous new structures that can sometimes lead to additional cyber-vulnerabilities.

Here we have tried to show the social, economic and competence losses triggered upon retirement can interact in the construction of a "new normal." For some well-resourced older adults, with good social networks, financial stability, and a range of post-retirement interests, the vulnerabilities are not so much tied to a paucity of resources, but may be associated with taking up new challenges. The retired doctor who lives alone and downloads the best-selling apps on a new smartphone has a different risk profile to the retired sales clerk who lives in close proximity to children and grandchildren and who is reliant on their second-hand devices and background knowledge. It is unfortunate that we don't fully recognise the different technology pathways possible following retirement, the way that these vary between individual, and the associated cyber-risks.

In terms of the implications of this work, particularly for policy and lifelong learning, we would argue the following. Firstly, we recognise that the workplace legacy knowledge for individuals will vary enormously. For those in manual labour, for example, the technology skills they possess upon retirement are unlikely to derive from workplace experience. But for those who do use technology in work, one policy recommendation we could make is to consider the extent to which, as a retirement offer, they could be given access to appropriate technical and cybersecurity expertise. On the approach to retirement, cybersecurity training packages could accompany existing retirement planning packages that are offered by some organisations. Naturally, this relies on the production of an effective cybersecurity training package that teaches the individual safe practices and where to find appropriate information. This provides challenges not only for policy makers, but also for researchers attempting to implement cybersecurity interventions targeted at older adults.

Secondly, additional support should be provided for those currently in retirement, provided in an accessible format way that empowers older adults to act safely online and promotes efficacy in engaging in safety behaviours. This is likely to begin with the promotion of government backed websites such as "Cyber Aware" in the United Kingdom, but should extend to provide an age appropriate source of information, which considers those with poorer computer literacy.

Finally, to addresses losses in day-to-day routine, social interaction, and feelings of sense of purpose, which may inadvertently lead to increased vulnerability, support should be provided to promote social groups for older adults that are empowered to provide cyber support, advice and guidance as well as provide a forum for support in which older adults can support each other. In this regard, recent work on the role of CyberGuardians within a support network is interesting.

Limitations and Future Work

We have discussed post-retirement losses without fully considering the interactions between these, noting that the interplay between these factors may intensify their effect on cybersecurity vulnerability. For example, an individual with limited financial and social resources may have to fall back on their own legacy knowledge – but what if they previously worked in a non-technical role with limited access to training? How does such an individual understand where to go to access good quality advice and support? Understanding the interplay of retirement factors is important in knowing how to target resources to support older adults.

In addition, we should consider more closely the way that cyber-attacks map onto the retirement transition. Oliveira et al. (2017) found that older adults are at particular risk of cyber-attacks associated with health, finances, and legal ideologies. Furthermore, attacks which involved reciprocation (an award was given and the email asked for recompense in the form of positive feedback) and social proofing (the incentive to join a holiday club with other similar adults) led to a significantly greater frequency of phishing link clicks. It is likely that retirees are particularly vulnerable to targeted attacks in domains that relate to their own particular retirement losses. For example, an individual in financial difficulties may be more likely to fall foul of financial phishing emails, and an individual who has lost a social network may be more likely to fall for holiday or romance scams that promise interaction with similar others. Preparing those approaching the retirement transition for the challenges they are likely to face and the associated threats may provide an interesting avenue for future cybersecurity interventions.

Finally, we have made a case for understanding more about heterogeneity in retirement, but we have done so on the basis of a study in which our sample of participants is not properly representative of our wider society. None of our participants were drawn from Black, Asian, Minority Ethnic (BAME) groups, none had declared disabilities. They were in relatively good health, many were homeowners and most were married or with partners. All of these factors are influential – but to take the last point as an example: the marital relationship influences *inter alia* post-retirement wellbeing (Szinovacz and Davey, 2003), leisure satisfaction (Losier et al., 1993), and decision as to when to retire (Smith and Moen, 1998). Additionally, it has also been implicated in specific cybersecurity risks such as an increased risk in consumer fraud victimisation in single older adults (Lee and Soberon-Ferrer, 1997).

Other issues to consider on this same issue of heterogeneity are type of retirement (phased or full) as this can influence retirement outcomes (de Vaus et al., 2007) and the time lapsed since retirement. There has been some debate around the duration of time it takes to "transition" fully into retirement (Reitzes and Mutran, 2004), but there is no doubt that the experiences given by retirees are likely to vary depending on the time since retiring. Recruitment criteria for this study required the participant to be within 5 years of their retirement and the main reason for this was to ensure that participant could remember the experiences of their retirement transition. Future research, especially any employing larger-scale quantitative methodology, might seek a broader age range of participants to pinpoint more of the socio-demographic issues but also to determine how the impact of type of retirement, former work type, and "drift" from the workplace can predict cybersecurity risk, behaviours, and/or attitudes.

CONCLUSION

This study sought to investigate how the retirement transition might lead to increased cyber-vulnerability in older adulthood. Through the use of one to one qualitative interviews with recently retired United Kingdom based older adults, we found that losses in social support structures, financial stability, and perceptions of declining competence can lead to changes in the way that technology is perceived and used. The changes to a retiree's technological landscape, in terms of both personal and external resources are likely to increase vulnerability to cyber-threats.

AUTHOR'S NOTE

This project contributes towards, but is not funded by, the EPSRC funded CSALSA (CyberSecurity Across the Life SpAn) Project in collaboration with Bath University, Portsmouth University, and Cranfield University (Grant Number: EP/P011454/1).

DATA AVAILABILITY STATEMENT

The datasets generated for this study will not be made publicly available as no quantitative data was collected, and the manuscripts may release person-identifiable information. Requests to access the datasets should be directed to the corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Northumbria University, Department of Psychology Ethics Committee. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

All authors have made substantial contributions to the conception or design of the work, and the acquisition, analysis, or interpretation of data for the work.

FUNDING

This study was supported by Northumbria University at Newcastle upon Tyne. This work was also supported by the Engineering and Physical Sciences Research Council (grant number EP/P011454/1).

REFERENCES

Age-UK (2015a). *Only the Tip of the Iceberg?: Fraud Against Older People–Evidence. Review*. London: Age-UK.

Age-UK (2015b). *Over Half of People Aged 65 + Targeted by Fraudsters. Age UK, 2015–2017*. Available online at: https://www.ageuk.org.uk/latest-press/archive/over-half-of-people-aged-65-targeted-by-fraudsters/# (accessed November, 2019).

Age-Uk (2018). *Financial Resilience During Retirement: Who is Well Placed to Cope with Life Events?*. Available online at: https://www.ageuk.org.uk/globalassets/age-uk/documents/reports-and-publications/reports-and-briefings/money-matters/rb_apr18_financial_resillience_summary.pdf (accessed November, 2019).

Ahvanooey, M. T., Li, Q., Rabbani, M., and Raza, A. (2017). A survey on smartphones security: software vulnerabilities, malware, and attacks. *Int. J. Adv. Comput. Sci. Appl.* 8, 30–45. doi: 10.14569/ijacsa.2017.081005

Alves, L., and Wilson, S. (2008). The effects of loneliness on telemarketing fraud vulnerability among older adults. *J. Elder Abuse Negl.* 20, 63–85. doi: 10.1300/J084v20n01

Barbosa, L. M., Monteiro, B., and Murta, S. G. (2016). Retirement adjustment predictors—a systematic review. *Work. Aging Retire.* 2, 262–280. doi: 10.1093/workar/waw008

Barnard, Y., Bradley, M. D., Hodgson, F., and Lloyd, A. D. (2013). Learning to use new technologies by older adults: Perceived difficulties, experimentation behaviour and usability. *Comput. Hum. Behav.* 29, 1715–1724. doi: 10.1016/j.chb.2013.02.006

Barnett, I., Guell, C., and Ogilvie, D. (2012). The experience of physical activity and the transition to retirement: a systematic review and integrative synthesis of qualitative and quantitative evidence. *Int. J. Behav. Nutr. Phys. Act.* 9:97. doi: 10.1186/1479-5868-9-97

Beck, F., Gillison, F., and Standage, M. (2010). A theoretical investigation of the development of physical activity habits in retirement. *Br. J. Health Psychol.* 15, 663–679. doi: 10.1348/135910709X479096

Bleidorn, W., and Schwaba, T. (2018). Retirement is associated with change in self-esteem. *Psychol. Aging* 33, 586–594. doi: 10.1037/pag0000253

Brooks, J., McCluskey, S., Turley, E., and King, N. (2015). The utility of template analysis in qualitative psychology research. *Qual. Res. Psychol.* 12, 202–222. doi: 10.1080/14780887.2014.955224

Buchanan, T., and Whitty, M. T. (2014). The online dating romance scam: causes and consequences of victimhood. *Psychol. Crime Law* 20, 261–283. doi: 10.1080/1068316X.2013.772180

Burr, A., Santo, J. B., and Pushkar, D. (2011). Affective well-being in retirement: the influence of values, money, and health across three years. *J. Happiness Stud.* 12, 17–40. doi: 10.1007/s10902-009-9173-2

Canetti, D., Gross, M., Waismel-Manor, I., Levanon, A., and Cohen, H. (2017). How cyberattacks terrorize: cortisol and personal insecurity jump in the wake of cyberattacks. *Cyberpsychol. Behav. Soc. Netw.* 20, 72–77. doi: 10.1089/cyber.2016.0338

Chakraborty, R., Vishik, C., and Rao, H. R. (2013). Privacy preserving actions of older adults on social media: exploring the behavior of opting out of information sharing. *Decis. Support Syst.* 55, 948–956. doi: 10.1016/j.dss.2013.01.004

Cho, J.-H., Cam, H., and Oltramari, A. (2016). "Effect of personality traits on trust and risk to phishing vulnerability: modeling and analysis," in *Proceedings of the International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support (CogSIMA)* (San Diego, CA: IEEE), 7–13. doi: 10.1109/COGSIMA.2016.7497779

Choi, N. G. (2001). Relationship between Life Satisfaction and Postretirement Employment among Older Women. *Int. J. Aging Hum. Dev.* 52, 45–70. doi: 10.2190/2W25-DH9H-2F4D-7HWX

Choi, K.-S. (2008). Cyber-routine activities. *Cyber Criminol.* 229–252. doi: 10.1201/b10718-19

Chopik, W. J. (2016). The benefits of social technology use among older adults are mediated by reduced loneliness. *Cyberpsychol. Behav. Soc. Netw.* 19, 551–556. doi: 10.1089/cyber.2016.0151

Damman, M., Henkens, K., and Kalmijn, M. (2015). Missing work after retirement: the role of life histories in the retirement adjustment process. *Gerontologist* 55, 802–813. doi: 10.1093/geront/gnt169

Davey, J. A. (2007). Older people and transport: coping without a car. *Ageing Soc.* 27, 49–65. doi: 10.1017/S0144686X06005332

de Vaus, D., Wells, Y., Kendig, H., and Quine, S. (2007). Does gradual retirement have better outcomes than abrupt retirement? Results from an Australian panel study. *Ageing Soc.* 27, 667–682. doi: 10.1017/S0144686X07006228

Dimond, J. P., Shehan Poole, E., and Yardi, S. (2010). "The effects of life disruptions on home technology routines," in *Proceedings of the 16th ACM International Conference on Supporting Group Work*, Vol. 10. (New York, NY: ACM), 85–88. doi: 10.1145/1880071.1880085

Dorfman, L. T. (1992). Academics and the transition to retirement. *Educ. Gerontol.* 18, 343–363. doi: 10.1080/0360127920180404

Durrant, A., Kirk, D., Trujillo Pisanty, D., Moncur, W., Orzech, K., Schofield, T., et al. (2017). "Transitions in digital personhood," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, Vol. 17, (New York, NY: Association for Computing Machinery), 6398–6411. doi: 10.1145/3025453.3025913

Ekerdt, D. J., and Koss, C. (2016). The task of time in retirement. *Ageing Soc.* 36, 1295–1311. doi: 10.1017/S0144686X15000367

Finkel, D., Andel, R., Gatz, M., and Pedersen, N. L. (2009). The role of occupational complexity in trajectories of cognitive aging before and after retirement. *Psychol. Aging* 24, 563–573. doi: 10.1037/a0015511

Forget, A., Pearman, S., Thomas, J., Acquisti, A., Christin, N., Cranor, L. F., et al. (2016). "Do or do not, there is no try: user engagement may not improve security outcomes," in *Proceedings of the Twelfth Symposium on Usable Privacy and Security (SOUPS), 97–111*, Denver, CO.

Fujs, D., Mihelič, A., and Vrhovec, S. L. R. (2019). "The power of interpretation: qualitative methods in cybersecurity research," in *Proceedings of the 14th International Conference on Availability, Reliability and Security (ARES '19)* (New York, NY: ACM), doi: 10.1145/3339252.3341479

Gordon, M. L., Gatys, L., Guestrin, C., Bigham, J. P., Trister, A., and Patel, K. (2019). "App usage predicts cognitive ability in older adults," in *Proceedings of CHI Conference on Human Factors in Computing Systems*, Glasgow, 1–12. doi: 10.1145/3290605.3300398

Grable, J. (2000). Financial risk tolerance and additional factors that affect risk taking in everyday money matters. *J. Bus. Psychol.* 14, 625–630. doi: 10.1023/A

Gregor, P., Newell, A., and Zajicek, M. (2002). "Designing for dynamic diversity – interfaces for older people," in *Proceedings of the ACM Conference on Assistive Technologies, ASSETS 2002*, Edinburgh, 151–156.

Grimes, G. A., Hough, M. G., Mazur, E., and Signorella, M. L. (2010). Older adults' knowledge of internet hazards. *Educ. Gerontol.* 36, 173–192. doi: 10.1080/03601270903183065

Grimes, G. A., Hough, M. G., and Signorella, M. L. (2007). Email end users and spam: relations of gender and age group to attitudes and actions. *Comput. Human Behav.* 23, 318–332. doi: 10.1016/j.chb.2004.10.015

Guan, J., and Huck, J. (2012). "Children in the digital age," in *Proceedings of the 2012 iConference*, (New York, NY: ACM), 506–507. doi: 10.1145/2132176.2132266

Guest, G., Bunce, A., and Johnson, L. (2006). How many interviews are enough: an experiment with data saturation and variability. *Field Methods* 18, 59–82. doi: 10.1177/1525822X05279903

Hauk, N., Hüffmeier, J., and Krumm, S. (2018). Ready to be a silver surfer? A meta-analysis on the relationship between chronological age and technology acceptance. *Comput. Hum. Behav.* 84, 304–319. doi: 10.1016/j.chb.2018.01.020

Henkens, K., Van Dalen, H. P., Ekerdt, D. J., Hershey, D. A., Hyde, M., Radl, J., et al. (2017). What we need to know about retirement: pressing issues for the coming decade. *Gerontologist* 58, 805–812. doi: 10.1093/geront/gnx095

Hill, R., Betts, L. R., and Gardner, S. E. (2015). Older adults experiences and perceptions of digital technology: (Dis)empowerment, wellbeing, and inclusion. *Comput. Hum. Behav.* 48, 415–423. doi: 10.1016/j.chb.2015.01.062

Hjorth, K., and Fosgerau, M. (2009). "Determinants of the degree of loss aversion," in *Paper Presented at the.International Choice Modelling Conference*. Harrogate. 1–25

Hutto, C., and Bell, C. (2014). "Social media gerontology: Understanding social media usage among a unique and expanding community of users," in *Proceedings of the 47th Hawaii International Conference on System Sciences* (Waikoloa, HI: IEEE), 1755–1764. doi: 10.1109/HICSS.2014.223

Jang-Jaccard, J., and Nepal, S. (2014). A survey of emerging threats in cybersecurity. *J. Comput. Syst. Sci.* 80, 973–993. doi: 10.1016/j.jcss.2014.02.005

Jeske, D., Briggs, P., and Coventry, L. (2016). Exploring the relationship between impulsivity and decision-making on mobile devices. *Pers. Ubiquitous Comput.* 20, 545–557. doi: 10.1007/s00779-016-0938-4

Jiang, M., Tsai, H., Yi, S., Cotten, S. R., Rifon, N. J., LaRose, R., et al. (2016). Generational differences in online safety perceptions, knowledge, and practices. *Educ. Gerontol* 42, 621–634. doi: 10.1080/03601277.2016.1205408

Juárez, M. A. R., González, V. M., and Favela, J. (2018). Effect of technology on aging perception. *Health Informatics J.* 24, 171–181. doi: 10.1177/1460458216661863

Khvorostianov, N., Elias, N., and Nimrod, G. (2012). "Without it I am nothing": the internet in the lives of older immigrants. *New Media Soc.* 14, 583–599. doi: 10.1177/1461444811421599

Kim, J. E., and Moen, P. (2002). Retirement transitions, gender, and psychological well-being a life-course, ecological model. *J. Gerontol. Ser. B Psychol. Sci. Soc. Sci.* 57, 212—-222. doi: 10.1093/geronb/57.3.P212

Kim, J. E., Moen, P., and Kim, J. E. (2001). Is retirement good or bad for subjective well-being? *Curr. Dir. Psychol. Sci.* 10, 83–86. doi: 10.1111/1467-8721.00121

King, N. (1998). *"Template Analysis.," in Qualitative Methods and Analysis in Organizational Research: A Practical Guide*. Thousand Oaks, CA: Sage Publications Ltd, 118–134.

Kirlappos, I., Parkin, S., and Sasse, M. A. (2014). Learning from "shadow security:" WHY UNDERSTANDING NON-COMPLIANT BEHAVIORS PROVIDES THE BASIS FOR EFFECTIVE SECURITY. *Papper Presended at the Workshop on Usable Security (USEC)*. (San Diego, CA: Internet Society), 2014–2016. doi: 10.14722/usec.2014.23007

Kisekka, V., Chakraborty, R., Bagchi-Sen, S., and Rao, H. R. (2015). Investigating factors influencing web-browsing safety efficacy (WSE) among older adults. *J. Inf. Priv. Secur.* 11, 158–173. doi: 10.1080/15536548.2015.1073534

Kloep, M., and Hendry, L. B. (2006). Pathways into retirement: entry or exit? *J. Occup. Organ. Psychol.* 79, 569–593. doi: 10.1348/096317905X68204

Lee, J., and Soberon-Ferrer, H. (1997). Consumer vulnerability to fraud: influencing factors. *J. Consum. Aff.* 31, 70–89. doi: 10.1111/j.1745-6606.1997.tb00827.x

Levitt, H. M., Bamberg, M., Creswell, J. W., Frost, D. M., Josselson, R., and Suárez-Orozco, C. (2018). Journal article reporting standards for qualitative primary, qualitative meta-analytic, and mixed methods research in psychology: the APA publications and communications board task force report. *Am. Psychol.* 73, 26–46. doi: 10.1037/amp0000151

Lindley, S. E., Harper, R., and Sellen, A. (2008). "Designing for elders: exploring the complexity of relationships in later life,'in *Proceedings of the 22nd Annual Conference*. New York, NY. 77–86.

Losier, G. F., Bourque, P. E., and Vallerand, R. J. (1993). A motivational model of leisure participation in the elderly. *J. Psychol. Interdiscip. Appl.* 127, 153–170. doi: 10.1080/00223980.1993.9915551

Lu, L., Li, Z., Wu, Z., Lee, W., and Jiang, G. (2012). "CHEX: statically vetting android apps for component hijacking vulnerabilities," in *Proceedings of the 2012 ACM Conference on Computer and Communications Security*, (New York, NY: ACM), 229–240. doi: 10.1145/2382196.2382223

Luiu, C., Tight, M., and Burrow, M. (2017). The unmet travel needs of the older population: a review of the literature. *Transp. Rev.* 37, 488–506. doi: 10.1080/01441647.2016.1252447

Mao, M., Blackwell, A. F., and Good, D. A. (2017). *Retirement Transition in the Digital Ecology: Reflecting on Identity Reconstruction and Technology Appropriation. ArXiv*. Available online at: http://arxiv.org/abs/1710.08867 (accessed November, 2019).

Marquié, J. C., Jourdan-Boddaert, L., and Huet, N. (2002). Do older adults underestimate their actual computer knowledge? *Behav. Inf. Technol.* 21, 273–280. doi: 10.1080/0144929021000020998

Martin, N., and Rice, J. (2013). Spearing high net wealth individuals. *Int. J. Inf. Secur. Priv.* 7, 1–15. doi: 10.4018/jisp.2013010101

Massimi, M., Dimond, J. P., and Le Dantec, C. A. (2012). "Finding a new normal: the role of technology in life disruptions," in *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, (New York, NY: ACM), 719–728.

Meng, A., Nexø, M. A., and Borg, V. (2017). The impact of retirement on age related cognitive decline – A systematic review. *BMC Geriatr.* 17:17. doi: 10.1186/s12877-017-0556-7

Mitzner, T., Boron, J., Fausset, C., Adams, A., Charness, N., Czaja, S., et al. (2010). Older adults talk technology: technology usage and attitudes tracy. *Comput. Hum. Behav. J.* 26, 1710–1721. doi: 10.1016/j.chb.2010.06.020

Nahum-Shani, I., and Bamberger, P. A. (2009). Work Hours, retirement and supportive relations among older adults. *J. Organ. Behav.* 30, 1–25. doi: 10.1021/nn300902w.Release

Nicholson, J., Coventry, L., and Briggs, P. (2013a). "Age-related performance issues for PIN and face-based authentication systems," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2013)* (New York, NY: ACM), 323–332. doi: 10.1145/2470654.2470701

Nicholson, J., Coventry, L., and Briggs, P. (2013b). Faces and Pictures: Understanding age differences in two types of graphical authentications. *Int. J. Hum. Comput. Stud.* 71, 958–966. doi: 10.1016/j.ijhcs.2013.07.001

Nicholson, J., Coventry, L., and Briggs, P. (2019). "If It's important it will be a headline: cybersecurity information seeking in older adults " if it ' s important it will be a headline ": cybersecurity information seeking in older adults," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (New York, NY: ACM), doi: 10.1145/3290605.3300579

Nowland, R., Necka, E. A., and Cacioppo, J. T. (2018). Loneliness and social internet use: pathways to reconnection in a digital world? *Perspect. Psychol. Sci.* 13, 70–87. doi: 10.1177/1745691617713052

Ofcom (2018). *Communications Market Report*. London: Ofcom.

Oliveira, D., Rocha, H., Yang, H., Ellis, D., Dommaraju, S., Muradoglu, M., et al. (2017). "Dissecting spear phishing emails: on the interplay of user age, weapons of influence, and life domains in predicting phishing susceptibility," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (New York, NY: ACM), doi: 10.1145/10.1145/3025453.302583

Olivier, S., Burls, T., Fenge, L. A., and Brown, K. (2015). "winning and losing": Vulnerability to mass marketing fraud. *J. Adult Prot.* 17, 360–370. doi: 10.1108/JAP-02-2015-0002

Orth, U., and Robins, R. W. (2014). The development of self-esteem. *Curr. Dir. Psychol. Sci.* 23, 381–387. doi: 10.1177/0963721414547414

Osborne, J. W. (2012). Psychological effects of the transition to retirement. *Can. J. Couns. Psychother.* 46, 45–58. doi: 10.1080/13642537.2012.734472

Parsons, K., Butavicius, M., Delfabbro, P., and Lillie, M. (2019). Predicting susceptibility to social influence in phishing emails. *Int. J. Hum. Comput. Stud.* 128, 17–26. doi: 10.1016/j.ijhcs.2019.02.007

Peek, S. T. M., Luijkx, K. G., Rijnaard, M. D., Nieboer, M. E., Van Der Voort, C. S., Aarts, S., et al. (2016). Older Adults' reasons for using technology while aging in place. *Gerontology* 62, 226–237. doi: 10.1159/000430949

Pettican, A., and Prior, S. (2011). Its a new way of life: An exploration of the occupational transition of retirement. *Br. J. Occup. Ther.* 74, 12–19. doi: 10.4276/030802211X12947686093521

Price, C. A. (2003). Professional women's retirement adjustment: the experience of reestablishing order. *J. Aging Stud.* 17, 341–355. doi: 10.1016/S0890-4065(03)00026-4

Rahman, N. A. A., Permatasari, F., and Hafsari, Y. (2017). A review on social media issues and security awareness among the users. *J. Appl. Technol. Innov* 1, 28–36.

Reitzes, D. C., and Mutran, E. J. (2004). The Transition to retirement: stages and factors that influence retirement adjustment. *Int. J. Aging Hum. Dev.* 59, 63–84. doi: 10.2190/NYPP-RFFP-5RFK-8EB8

Robertson, D. A., and Kenny, R. A. (2016). Negative perceptions of aging modify the association between frailty and cognitive function in older adults. *Pers. Individ. Dif.* 100, 120–125. doi: 10.1016/j.paid.2015.12.010

Saini, H., Rao, Y. S., and Panda, T. C. (2012). Cyber-Crimes and their Impacts: a Review. *Int. J. Eng. Res. Appl.* 2, 202–209.

Salovaara, A., Lehmuskallio, A., Hedman, L., Valkonen, P., and Nasanen, J. (2010). Information technologies and transitions in the lives of 55-65-year-olds: the case of colliding life interests. *Int. J. Hum. Comput. Stud.* 68, 803–821. doi: 10.1016/j.ijhcs.2010.06.007

Salthouse, T. A. (2009). When does age-related cognitive decline begin? *Neurobiol. Aging* 30, 507–514. doi: 10.1016/j.neurobiolaging.2008.09.023

Sargent-Cox, K. A., Anstey, K. J., and Luszcz, M. A. (2012). The relationship between change in self-perceptions of aging and physical functioning in older adults. *Psychol. Aging* 27, 750–760. doi: 10.1037/a0027578

Sarno, D. M., Lewis, J. E., Bohil, C. J., Shoss, M. K., and Neider, M. B. (2017). "Who are phishers luring: a demographic analysis of those susceptible to fake emails," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. Thousand Oaks, CA: SAGE. doi: 10.1177/1541931213601915

Seeman, T. E., Unger, J. B., Mcavay, G. M., and de, L. C. (1999). Self-Efficacy beliefs and perceived declines in functional ability: mc arthurs studies of seccessful aging. *J. Gerontol.* 54, 214–222.

Seifert, A., and Schelling, H. R. (2018). Seniors online: Attitudes toward the internet and coping with everyday life. *J. Appl. Gerontol.* 37, 99–109. doi: 10.1177/0733464816669805

Selwyn, N. (2004). The information aged: A qualitative study of older adults' use of information and communications technology. *J. Aging Stud.* 18, 369–384. doi: 10.1016/j.jaging.2004.06.008

Shao, J., Zhang, Q., Ren, Y., Li, X., and Lin, T. (2019). Why are older adults victims of fraud? Current knowledge and prospects regarding older adults' vulnerability to fraud. *J. Elder Abus. Negl.* 31, 225–243. doi: 10.1080/08946566.2019.1625842

Sheehy-skeffington, J., and Rea, J. (2017). *How Poverty affects People's Decision-Making Processes. Www.Jrf.Org.Uk*, 1–73. Available online at: https://www.jrf.org.uk/report/how-poverty-affects-peoples-decision-making-processes (accessed November, 2019).

Shultz, K. S., and Wang, M. (2011). Psychological perspectives on the changing nature of retirement. *Am. Psychol.* 66, 170–179. doi: 10.1037/a0022411

Siegenthaler, K. L., and Vaughan, J. (1998). Older women in retirement communities: perceptions of recreation and leisure. *Leis. Sci.* 20, 53–66. doi: 10.1080/01490409809512264

Smith, D. B., and Moen, P. (1998). Spousal influence on retirement: his, her, and their perceptions. *J. Marriage Fam.* 60, 734. doi: 10.2307/353542

Szinovacz, M. E., and Davey, A. (2003). Honeymoons and joint luncheons: effects of spouse's employment on depressive symptoms. *Conf. Pap. Am. Sociol. Assoc.* 59, 1–20.

Thoits, P. A. (2012). Role-identity salience, purpose and meaning in life, and well-being among volunteers. *Soc. Psychol. Q.* 75, 360–384. doi: 10.1177/0190272512459662

Torrens-Burton, A., Basoudan, N., Bayer, A. J., and Tales, A. (2017). Perception and reality of cognitive function: information processing speed, perceived memory function, and perceived task difficulty in older adults. *J. Alzheimes Dis.* 60, 1601–1609. doi: 10.3233/JAD-170599

Tosun, L. P. (2012). Motives for facebook use and expressing "true self" on the internet. *Comput. Hum. Behav.* 28, 1510–1517. doi: 10.1016/j.chb.2012.03.018

van Solinge, H., and Henkens, K. (2008). Adjustment to and satisfaction with retirement: two of a kind? *Psychol. Aging* 23, 422–434. doi: 10.1037/0882-7974.23.2.422

Vaportzis, E., Clausen, M. G., and Gow, A. J. (2017). Older adults perceptions of technology and barriers to interacting with tablet computers: a focus group study. *Front. Psychol.* 8:01687. doi: 10.3389/fpsyg.2017.01687

Vines, J., Pritchard, G., Wright, P., and Olivier, P. (2015). An age-old problem: examining the discourses of ageing in hci and strategies for future research. *Tochi* 22, 1–27. doi: 10.1145/2696867

Von Solms, R., and Van Niekerk, J. (2013). From information security to cyber security. *Comput. Secur.* 38, 97–102. doi: 10.1016/j.cose.2013.04.004

Wang, M. (2007). Profiling retirees in the retirement transition and adjustment process: examining the longitudinal change patterns of retirees' psychological well-being. *J. Appl. Psychol.* 92, 455–474. doi: 10.1037/0021-9010.92. 2.455

Wang, M., Henkens, K., and van Solinge, H. (2011). Retirement adjustment: a review of theoretical and empirical advancements. *Am. Psychol.* 66, 204–213. doi: 10.1037/a0022414

Whitty, M., Doodson, J., Creese, S., and Hodges, D. (2015). Individual differences in cyber security behaviors: an examination of who is sharing passwords. *Cyberpsychol Behav. Soc. Netw.* 18, 3–7. doi: 10.1089/cyber.2014.0179

Whitty, M. T. (2017). Do you love me? psychological characteristics of romance scam victims. *Cyberpsychol. Behav. Soc Netw.* 21, 105–109. doi: 10.1089/cyber.2016.0729

Wittes, B., Poplin, C., Jurecic, Q., and Spera, C. (2016). *Sextortion: Cybersecurity, Teenagers, and Remote Sexual Assault 1. Cent. Tecnol. Innov. BROOKINGS*, 1–47. Available online at: https://www.brookings.edu/research/sextortion-cybersecurity-teenagers-and-remote-sexual-assault/ (accessed November, 2019).

Workman, M., Bommer, W. H., and Straub, D. (2008). Security lapses and the omission of information security measures: a threat control model and empirical test. *Comput. Human Behav.* 24, 2799–2816. doi: 10.1016/j.chb.2008.04.005

Young, A., and Tinker, A. (2017). Who are the baby boomers of the 1960s? *Work. with Older People* 21, 197–205. doi: 10.1108/WWOP-06-2017-0015 doi: 10.1108/wwop-06-2017-0015

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 05 May 2020
doi: 10.3389/fpsyg.2020.00840

Assessing the Factors Associated With the Detection of Juvenile Hacking Behaviors

*Jin Ree Lee and Thomas J. Holt**

School of Criminal Justice, College of Social Science, Michigan State University, East Lansing, MI, United States

Research on delinquency reduction often highlights the importance of identifying and sanctioning antisocial and illegal activities so as to reduce the likelihood of future offending. The rise of digital technology complicates the process of detecting cybercrimes and technology enabled offenses, as individuals can use devices from anywhere to engage in various harmful activities that may appear benign to an observer. Despite the growth of cybercrime research, limited studies have examined the extent to which technology enabled offenses are detected, or the behavioral and attitudinal factors associated with being unobserved or caught for one's actions. The current study addresses this gap in the literature by estimating a multinomial regression model for self-reported computer hacking behavior and the likelihood of those actions being detected in a large international sample of juveniles (*N* = 51,059). The findings demonstrate significant differences between youth who hack without detection compared to those who are caught. The implications of this analysis for our understanding of cybercrime and its relationship to traditional delinquency are explored in depth.

Keywords: computer hacking, low self-control, social bonds, cybercrime, deterrence, juvenile delinquency

OPEN ACCESS

Edited by:
Paul Watters,
La Trobe University, Australia

Reviewed by:
Stephen McCombie,
Macquarie University, Australia
Mamoun Alazab,
Charles Darwin University, Australia

****Correspondence:***
Thomas J. Holt
holtt@msu.edu

Specialty section:
This article was submitted to Cognition, a section of the journal Frontiers in Psychology

Received: *10 January 2020*
Accepted: *06 April 2020*
Published: *05 May 2020*

Citation:
Lee JR and Holt TJ (2020) Assessing the Factors Associated With the Detection of Juvenile Hacking Behaviors. Front. Psychol. 11:840. doi: 10.3389/fpsyg.2020.00840

ASSESSING THE FACTORS ASSOCIATED WITH THE DETECTION OF JUVENILE HACKING BEHAVIORS

For many forms of crime and delinquency, the notion of deterring behavior is imperative so as to reduce the risk of future offending. Deterrence is generally derived from the perceived threat of detection and sanctioning for wrongdoing, whether from police or informal sources of control such as peers or parents in the case of delinquency (Nagin and Pogarsky, 2001; Pratt et al., 2006). The decision to offend is thus a calculus of the perceived likelihood of detection relative to the reward acquired from the offense (Cornish and Clarke, 2014). Detection is, however, variable based on the nature of the offense and its situational characteristics, such as the presence of surveillance tools and observers to report wrongdoing (Clarke, 1997; Cornish and Clarke, 2003; Reyns, 2010). As a result, the risk of detection varies based on the extent to which offenders can obfuscate their behaviors and otherwise appear to engage in normal behaviors in physical space (Wright and Decker, 1996; Cherbonneau and Copes, 2005; Cardone and Hayes, 2012).

The rise of computers and the Internet have created new opportunities to engage in crimes that are more difficult to detect through traditional means (Yar, 2013; Holt and Bossler, 2016). Individuals can engage in so-called cybercrimes where their use of technology enables them to commit an offense from the comfort of their home without the need to interact with their victims in public settings (Holt and Bossler, 2016). In addition, parents and/or guardians who may observe offline deviant behaviors may not notice cybercriminality because the individual may simply appear to be typing on a keyboard or utilizing a specific program to access content (Holt and Bossler, 2016). Actors may also conceal illegal online activitiy by taking their laptop or electronic device into a private space so as to avoid being asked questions by family members or guardians (Holt et al., 2019).

These factors may all lower the perceived risk of detection for engaging in cybercrime, as the rate of arrest is extremely low proportionally to physical crimes (see Holt and Bossler, 2016). This is especially true for computer hacking, generally defined as the use of technological understanding to engage in unauthorized access of computer systems and networks (Jordan and Taylor, 1998; Wall, 2001; Furnell, 2002; Schell and Dodge, 2002; Holt, 2007; Holt et al., 2019). Though hacking can be used for legitimate applications, the behavior has largely been associated with malicious, criminal activity in the general public over the last two decades (Furnell, 2002; Holt, 2007; Grabosky, 2016). As a result, hacking is frequently viewed as a serious threat affecting both the public and private sector.

Research regarding hackers and hacking have increased over the last two decades, providing insight into key individual predictors for hacking among juvenile and adult samples (see Holt and Bossler, 2016 for review). Research examining the detection of hacking is nascent in the broader literature (see Maimon et al., 2014), calling to question what factors differentiate hackers from non-hackers as to their likelihood of being caught for their involvement in an increasingly common form of cybercrime. Such information is vital to better understand the factors that may increase an actor's willingness to hack, as well as decrease their likelihood of detection. In turn, better detection and prevention strategies can be developed to curb hacking behavior among youth, regardless of their ability to conceal their actions.

The current study attempted to address this question through the use of a clustered multinomial regression model of an international sample of over 51,000 juveniles. The model compared those who hacked and avoided detection as well as those who were detected, against the larger sample of youth who did not hack. The findings demonstrated key differences in the behaviors and attitudes of youth on the basis of their risk of detection, particularly regarding their access to technology and levels of parental supervision. The implications of this analysis for our understanding of ways to deter early onset hacking, and hacker behavior more generally, were discussed in detail.

UNDERSTANDING COMPUTER HACKING AND HACKER BEHAVIORS

Social science research over the last few decades have revealed hacking to be a skill set that can be applied for both malicious and/or legitimate purposes (Holt, 2007, 2010; Holt et al., 2010; Steinmetz, 2015, 2017). The concept of hacking emerged in the 1950s at the Massachusetts Institute of Technology as a way to reference the manipulation of technology to produce an outcome that was different form its intended use (Levy, 1984). Hacking as a form of non-deviant manipulation has continued through today, including open-source software programming and computer hardware manipulation (Levy, 1984; Taylor, 1999; Coleman, 2014).

At the same time, a proportion of individuals engage in hacking for criminal applications, affecting business, citizens, and governments (Steinmetz, 2015). The rise of criminal hacking began in the late 1970s and 1980s, concurrent with the growth of personal computers and rudimentary Internet connectivity (e.g., Hollinger and Lanza-Kaduce, 1988). Juveniles became interested in technology during this period, using their expertise to hack financial systems and sensitive networks (Slatalla and Quittner, 1995; Furnell, 2002; Schell and Dodge, 2002). In fact, small groups of teenage hackers with names like the "414 gang" and the "Masters of Deception" targeted high-profile companies and infrastructure, generating significant concern over the way youth may become involved in criminal activities online (Slatalla and Quittner, 1995; Calce and Silverman, 2008; Yar, 2013).

Qualitative research has found that the onset of hacking occurs during early adolescence, similar to offline forms of anti-social and deviant behavior (Jordan and Taylor, 1998; Holt, 2007). During this period, individuals tend to engage in minor, simplistic hacks as they gain insight into computer technology and methods of hacking generally (Taylor, 1999; Holt, 2007). As one's technical skill increases, so does the escalation of their offending frequency and severity. As a result, there is a need to understand the factors associated with the detection of hacking during this period so as to improve our comprehension of potential desistance factors that may reduce hacking over the long term (Maimon et al., 2014; Holt and Bossler, 2016; NCA, 2017).

Few studies have considered the factors that may be associated with the detection of hacking during adolescence, or during late adolescence in college samples (see Maimon et al., 2014; Holt and Bossler, 2016). Traditional criminological theories provide direction for factors that may be associated with an increased likelihood of being caught engaging in delinquent behaviors, including hacking. In fact, multiple correlates of hacking are consistent with predictors of traditional acts of crime and delinquency. To that end, Gottfredson and Hirschi's (1990) general theory of crime has been found to predict individual involvement in hacking behaviors, such as password guessing to access accounts and alter content without permission from the owner (Bossler and Burruss, 2011; Holt et al., 2012, 2019; Marcum et al., 2014; Udris, 2016). Gottfredson and Hirschi (1990) argued that crime is a choice derived from weighing the costs and benefits of offending, including the risk of detection. They suggest this decision is influenced by one's level of self-control when presented with opportunities to offend.

The level of self-control an individual has is a result of their parents' ability to monitor, recognize, and punish deviant behavior when they occur, thereby instilling a capacity to regulate one's actions in the moment (Gottfredson and Hirschi, 1990). Self-control is also established in early childhood, possibly accounting for early onset delinquent and anti-social behaviors (Pratt and Cullen, 2000; Vazsonyi et al., 2017). In essence, individuals with higher levels of self-control are more likely to restrain themselves when encountering criminal opportunities, while those with lower levels of self-control are more likely to take advantage of those same opportunities even when higher levels of risk detection are present. As a result, it is hypothesized that youth who hack are more likely to have low self-control compared to the general population, regardless of their risk of detection.

In much the same way, involvement in hacking is situationally dependent on access to computers and Internet connectivity. The role of opportunity as a predictor for hacking is under-examined, however, especially among juvenile populations (see Holt et al., 2019). To that end, it is virtually impossible to hack without having access to computers, mobile devices, and the Internet. Technology is readily available in most nations, creating near-constant opportunities to offend. As a consequence, criminological research demonstrates an important association between factors that increase the perceived risk and effort involved in committing an offense and reduced willingness to act on criminal opportunities (Cohen and Felson, 1979; Felson, 1986, 1995; Reyns, 2010). Resources that increase behavioral monitoring and create opportunities to intervene in offending activities may reduce individuals' situational willingness to offend (Reyns, 2010).

Various studies have examined opportunity factors and cybercrime offending with varying results. For one, Maimon et al. (2014) investigated the influence of a warning banner on the frequency and duration of hacking incidents directed at computer systems online (see also Wilson et al., 2015). The study found that while the use of warning banners did not lead to an immediate discontinuation of the hacking incident, it reduced the duration of each hacking incident. These findings support the proposition that increased risk of detection may decrease the offending behaviors of motivated hackers.

Since many individuals report engaging in early hacking behaviors at home (Taylor, 1999; Holt, 2007), increased monitoring of computer use or limiting the amount of time one spends on the computer may reduce opportunities to hack. Similarly, the more supervision and monitoring of computer activity, the more likely an individual's actions will be observed and punished (Marcum et al., 2014). There may, however, be economic barriers to technology access that may affect an individual's risk of detection for hacking. Families that only own one computer may keep it in an open place where it can be accessed by all, making its usage more easily observed by parents and/or guardians. In contrast, youth who own their own device may be able to conceal their actions from others more easily. Similarly, youth who have their own rooms may encounter lower levels of detection from parents and/or guardians (e.g., parental supervision), as technology use is harder to monitor and supervise in closed areas than in open spaces.

An additional element that may be associated with hacking and the risk of detection is youths' relationship with their parents and/or guardians. Research has found a consistent relationship between parental bonds and delinquency, as those with weak attachments to parents are at greater risk of engaging in deviance (Hirschi, 1969; Gottfredson and Hirschi, 1990; Sampson and Laub, 2003). Further, a lack of emotional ties to one's parents may diminish their capacity to regulate behavior, negatively impacting their capacity to form relationships with pro-social peers throughout adolescence (Wright et al., 1999; Li, 2004). Parental supervision is also an important element to detecting delinquent and anti-social behaviors in the home, as noted across multiple criminological theories (Hirschi, 1969; Gottfredson and Hirschi, 1990; Sampson and Laub, 2003). When parents are able to exert direct control over their children through behavioral monitoring and punishing anti-social behavior, they are more likely to reduce their child's involvement in delinquency (Sampson and Laub, 2003).

The role of parental bonds with respect to hacking is particularly salient as individuals are most likely to hack while at home due to ease of access to computers, and greater uninterrupted time while using the device. Limited research revealed a significant association between strong social bonds, high self-control, and reduced risk of hacking among Korean youth (Kong and Lim, 2012; Bae, 2017). Similarly, two recent studies utilizing an international population of juveniles found a relationship between reduced parental supervision, low self-control, and self-reported hacking (Udris, 2016; Back et al., 2018). It is hypothesized that those with weaker parental attachment and lower parental supervision will be more likely to hack without being detected. In contrast, those who are detected will likely have weaker parental attachments and higher levels of supervision, increasing their risk of detection.

There are also demographic factors that may shape the risk of both involvement in hacking and the likelihood of detection. First, there is a clear gender difference in the rates of hacking reported in both quantitative and qualitative samples (Gilboa, 1996; Jordan and Taylor, 1998; Taylor, 1999; Schell and Dodge, 2002; Hutchings and Chua, 2017; Holt et al., 2019). Males report higher levels of hacking, which appears to be a result of differential access to technology between the sexes from early ages (Taylor, 1999; Hutchings and Chua, 2017). There is less research considering whether girls who hack are more likely to be detected than boys at early ages. Evidence suggests females may experience greater levels of parental supervision which reduce available opportunities to offend, even in online spaces (Daigle et al., 2007; Lanctôt and Guay, 2014). As a result, boys may be more likely to hack though there may be no gender difference with respect to the risk of detection for hacking.

A small number of studies also demonstrate that individuals who hack may be from higher socio-economic status backgrounds and larger cities due to greater access to technology (Schell and Dodge, 2002; Holt, 2007; Steinmetz, 2016). Few quantitative studies have examined this relationship (Marcum et al., 2014; Holt and Bossler, 2016), though recent research by Holt et al. (2019) found that youth in smaller cities and higher socio-economic status families were more likely to self-report hacking during adolescence. It may be that families in higher socio-economic status groups provide opportunities for technology use, which reduces their risk of detection. Similarly, youth living in smaller cities may have an increased risk of detection because of reduced opportunities for unstructured socialization, as well as greater social bonds to parents (Gardner and Shoemaker, 1989).

The Current Study

Though our understanding of hacking has increased substantially over the last two decades, research assessing the factors that predict an individual's involvement and detection in hacking are scant. This study tested multiple hypotheses regarding the risk of detection for computer hacking which has been largely under-examined in social sciene research to date (Maimon et al., 2014; Holt and Bossler, 2016). First, it is expected that individuals with low self-control will be more likely to hack, regardless of the likelihood of detection (e.g., Holt et al., 2012; Marcum et al., 2014; Udris, 2016; Back et al., 2018). Second, youth with greater access to and frequent use of technology in private settings will be more likely to hack without detection. Third, those who engage in piracy and spend more time with peers will be more likely to hack overall, regardless of their likelihood of detection. Fourth, youth with weaker parental attachments and supervision will be more likely to hack without being detected, though those who are detected will have no difference from the general population in terms of their level of supervision.

Fifth, socio-economic status may be associated with hacking and reduced risk of detection because of greater access to technology. Sixth, geographic location may influence the risk of detection for those living in smaller towns due to differences in parental monitoring and bonding. Finally, it is expected that males will be more likely to report hacking behaviors regardless of their risk of detection due to the gendered nature of hacking. The implications of this analysis for our understanding of the factors affecting individuals' risk of detection, as well as effective prevention and intervention efforts to affect juvenile hacking, were discussed in detail.

Data and Methods

To test the proposed hypotheses, this analysis utilized the Second International Self-Report of Delinquency study dataset (ISRD-2, Junger-Tas, 2010; Junger-Tas and Marshall, 2012). The respondent population of the ISRD-2 consisted of juveniles in grades 7 through 9 across 30 nations, representing North America, Latin America, and some of the EU.[1] Probability sampling was used in classrooms nested within schools to obtain respondents in small and large cities within each country (see Marshall and Enzmann, 2012 for more detail). Surveys were administered between 2005 and 2007 in school classrooms for students to complete via pencil and paper instruments. Computerized questionnaires were administered in Denmark, Finland, and Switzerland, though the data is not different from that of the larger survey population. Additionally, the sample included students attending public, private, vocational, and technical schools to reflect the diversity of educational experiences.

Such a dataset is essential to examine the extent to which hacking behaviors are identified among those who hack, as this question has yet to be addressed in survey research to date (Holt and Bossler, 2016). Furthermore, there is generally little research cultivating international samples of youth to assess their self-reported hacking behaviors (Taylor, 1999; Holt et al., 2019). The ISRD-2 is one of the few data sets available that provides a sufficient population to identify any behavioral, attitudinal, and demographic correlates of hacking behaviors and the risk of detection for these activities.

The full dataset contained 68,507 respondents, however, the final sample used in this analysis consisted of 51,059 based on missing or incomplete data. The loss of 25% of the total population did not affect the representative nature of the sample, as the respondent population resembled the original data set with respect to gender (49.2% female and 48.8% male) and age (mean = 1.08 in both samples). Additionally, the data were relatively equal with regard to geographic distribution: 26.8% of the final sample livedd in cities with less than 100,000 residents compared to 22.4% in the overall sample.

Dependent Variable

The dependent variable for the current study was juveniles' self-reported involvement in hacking. Respondents were asked if they ever used a computer for "hacking," and to specify if "the last time you did it were you found out?" A relatively small proportion of respondents reported engaging in hacking behaviors at any time (N = 3,733; 7.3%), and only 25.2% of those individuals (N = 943) were detected (see **Table 1**). Though the overall rate of self-reported hacking is relatively low, it is consistent with prior rates reported among youth (Holt et al., 2012; Marcum et al., 2014) and late adolescent populations (Skinner and Fream, 1997; Rogers et al., 2006; Bossler and Burruss, 2011; Holt et al., 2010). The relatively small number of individuals who reported being detected for hacking allowed for the construction of a three-item variable: those who did not hack (0), those who hacked and were not discovered (1), and those who hacked and were caught (2). This measure enabled a comparison between those who did not report hacking against the other two categories which reflected 5.5% and 1.8% of the sample respectively.

It is important to note that the measure used in this survey did not define what constitutes hacking, which is different from the broader quantitative literature on hacking (Bossler and Burruss, 2011; Holt et al., 2012; Marcum et al., 2014). This measure does not enable an assessment of specific factors unique to any form of hacking that may have increased the risk of detection, such as the target of the offense or the technical skills needed to complete the activity (Holt, 2007; Steinmetz, 2016). At the same time, the use of a more general measure allowed respondents to identify what they considered as a hack without any value judgments as to whether the hack was legitimate or unethical (Holt, 2007; Steinmetz, 2016). This sort of measure may be more reflective of the diverse range of behaviors associated with hacking, including both minor and serious activities as well as those with ethical and malicious applications (Jordan and Taylor, 1998; Taylor, 1999; Holt, 2007; Steinmetz, 2016).

Independent Variables

To assess opportunities to use technology, two binary variables were created from the following items: 1) "Do you have a computer at home that you are allowed to use?" (*own computer*) and 2) "Do you own a mobile phone?" (*own mobile*). A third opportunity measure was included to assess the impact of having

TABLE 1 | Descriptive Statistics (*N* = 51,059), Clustered by School (*N* = 1,183).

Variables	Description	N	Mean	SD	Min.	Max.
Dependent Variables						
Hacking Behavior			0.092	0.346	0	2
	0 = Did not hack	27,325				
	1 = Hacked/not detected	2,790				
	2 = Hacked/detected	943				
Opportunity Characteristics						
Own Computer			0.854	0.351	0	1
	0 = No	7,478				
	1 = Yes	43,581				
Own Mobile			0.896	0.303	0	1
	0 = No	5,310				
	1 = Yes	45,749				
Own Room			0.754	0.429	0	1
	0 = No	12,535				
	1 = Yes	38,524				
Engagement Characteristics						
Technology Use			4.184	1.363	1	6
	1 = None	1,358				
	2 = 1/2 h	4,270				
	3 = 1 h	10,993				
	4 = 2 h	13,236				
	5 = 3 h	9,401				
	6 = 4 h +	11,801				
Piracy			0.490	0.499	0	1
	0 = No	26,042				
	1 = Yes	25,017				
Contextual Characteristics						
Self-Control	12-item additive index, α = 0.83		60.674	20.252	0	100
Family Bond	4-item additive index, α = 0.55		80.636	17.0566	0	100
Time Peers			4.212	1.653	1	6
	1 = None	5,011				
	2 = 1/2 h	3,893				
	3 = 1 h	7,651				
	4 = 2 h	9,451				
	5 = 3 h	8,793				
	6 = 4 h +	16,260				
Parental Supervision			2.556	0.590	1	3
	1 = Never	2,622				
	2 = Sometimes	17,428				
	3 = Always	31,009				
Demographic Characteristics						
Age			1.089	0.272	0	3
	0 = Less than 12	49				
	1 = 12 to 15	47,161				
	2 = 16–17	3,655				
	3 = 18 and older	102				
Gender			0.489	0.499	0	1
	0 = Female	26,111				
	1 = Male	24,948				

(Continued)

TABLE 1 | Continued

Variables	Description	N	Mean	SD	Min.	Max.
Car Ownership			0.874	0.330	0	1
	0 = No	6,451				
	1 = Yes	44,608				
Small City			0.268	0.443	0	1
	0 = Larger than 100,000	37,375				
	1 = Smaller than 100,000	13,684				

TABLE 2 | Multinomial Regression Model for Hacking and Detection (*N* = 51,059), Clustered by School (*N* = 1,183).

	Hacked/Not Detected		Hacked/Detected	
Variables	**Coef.**	**SE**	**Coef.**	**SE**
Opportunity Characteristics				
Own Computer (1 = Yes)	0.335***	0.084	0.158	0.129
Own Mobile (1 = Yes)	0.211*	0.087	−0.021	0.128
Own Room (1 = Yes)	0.082	0.053	0.100	0.086
Engagement Characteristics				
Technology Use	0.144***	0.018	0.027	0.028
Piracy (1 = Yes)	1.618***	0.058	1.508***	0.090
Contextual Characteristics				
Self-Control	−0.016***	0.001	−0.013***	0.002
Family Bond	−0.006***	0.001	−0.007***	0.002
Time Peers	0.045**	0.014	0.063**	0.023
Parental Supervision	−0.293***	0.034	−0.019	0.059
Demographic Characteristics				
Age	0.065	0.036	0.007	0.078
Gender (1 = Male)	1.158***	0.047	0.823***	0.073
Car Ownership	0.181*	0.076	0.209	0.127
Small City	0.076	0.046	0.657***	0.067

*F = 122.60***; *p < 0.05, **p < 0.01, ***p < 0.001. [1]This study was conducted in 15 western European countries (Austria, Belgium, Denmark, Finland, France, Germany, Ireland, Italy, The Netherlands, Portugal, Spain, Sweden, Iceland, Norway, and Switzerland), 10 eastern European countries (Cyprus, the Czech Republic, Estonia, Hungary, Lithuania, Poland, Slovenia, Armenia, Bosnia-Herzegovina, and Russia), the United States (Illinois, Massachusetts, New Hampshire and Texas), and several countries outside of Europe and North America (Aruba, Netherlands Antilles, Suriname, and Venezuela).*

a personal space where an individual may be able to utilize a computer: "do you have a room of your own?" (*own room*: 0 = no; 1 = yes).

A set of two measures were included to examine the relationship technology use and online activities. One item assessed: 1) "Outside school how much time do you spend on an average school day on each of these activities: watching tv, playing games, or chatting on the computer?" using a six-item response: (*tech use*: 1 = "none"; 2 = "30 min"; 3 = "one hour"; 4 = "two hours"; 5 = "three hours"; 6 = "four hours plus"). The second item captured individual's self-reported digital piracy through responses to the following question: "when you use a computer did you ever download music or films during the last 12 months?" (*piracy*: 0 = no; 1 = yes).

To measure *self-control,* a variable was created using responses to 12 of the original 24-item index created by Grasmick et al. (1993). The measures capture four of the six dimensions of self-control (i.e., impulsivity, risk-taking, volatile temperament, and self-centeredness), which has been validated through prior research (Marshall and Enzmann, 2012; Rocque et al., 2013; Botchkovar et al., 2015). The Percentage of Maximum Possible (POMP) scoring method was used to create the measure for this analysis by first rescaling the 12-item measures from 0 to 100 to create an average score for each respondent (alpha = 0.83). Lower individual scores reflected lower levels of self-control.

In order to assess the relationship between time spent with peers, hacking, and the likelihood of detection, a six-item measure was created based on responses to the question: "Outside school how much time do you spend on an average school day.hanging out with friends" (*time peers*: 1 = "none"; 2 = "30 min"; 3 = "one hour"; 4 = "two hours"; 5 = "three hours"; 6 = "four hours plus"). It is hypothesized that increased time spent with peers should increase opportunities to offend, whether on or off-line (Osgood et al., 1996; Haynie and Osgood, 2005; Hoeben et al., 2016).

To measure family bonding, a mean score was created from the following four items: (1) "how do you usually get along with the man you live with (father, stepfather...)"; (2) "how do you usually get along with the woman you live with (your mother or stepmother)?"; (3) "how often do you and your parents (or the adults you live with) do something together, such as going to the movies, going or a walk or hike, visiting relatives, attending a sporting event, and things like that?"; and (4) "how many days a week do you usually eat the evening meal with (one of) your parents (or the adults you live with)?" Responses for each item were summed and then transformed into a POMP measure (Cohen et al., 1999), with higher scores indicating greater presence of the measure. The use of this POMP family bonding scale is common in studies utilizing the ISRD-2 (Botchkovar et al., 2015; Posick and Rocque, 2015) to produce a reliable family bonding measure (alpha = 0.55). An additional measure of parental supervision of general behavior was also included, asking respondents: "Do your parents (or the adults you live with) usually know who you are with when you go out?" A three-item response was provided (*parsup*: 1 = "never"; 2 = "sometimes"; 3 = "always/I don't go out").

To examine the hypotheses related to demographic factors and hacking, a set of four measures were used in this analysis.

A four-item measure for *age* was included (0 = "less than 12"; 1 = "12 to 15"; 2 = "16–17"; 3 = "18 and older"), along with a binary measure for family *car ownership* (0 = no; 1 = yes) as a proxy for both socio-economic status (see also Holt et al., 2019). A binary measure was included to capture whether the respondent lived in an important city within their country, or a large city or town of more than 100,000 residents, or one with less than 100,000 or not considered important relative to their nation (*small city*: 0 = more than 100,000; 1 = less than 100,000). Lastly, a binary measure was created for gender (0 = female; 1 = male) to examine its relationship to self-reported hacking and likelihood of detection (Bachmann, 2010; Gilboa, 1996; Hutchings and Chua, 2017).

Findings

To assess the behavioral and attitudinal factors associated with hacking and the risk of detection, a multinomial regression model was estimated (see **Table 2**). Respondents who did not self-report involvement in hacking within the last year served as the reference category, compared to those who hacked without detection, and those who hacked and were caught. The large number of respondents across the various countries sampled created unique variations within and across the study populations. The regressions were estimated using the cluster command by school (N = 1,183) using STATA-13 statistical software to reduce the size of both the intra-cluster correlations and standard errors. No evidence of multicollinearity could be found between the variables in the models, as no VIF was higher than 1.22, while no tolerance was lower than 0.81. The findings demonstrated key differences between these populations. First, those who hacked without detection were more likely to have their own computer and mobile device than those who did not hack. Additionally, they were more likely to spend greater amounts of time on a computer or television, as well as spend more time with peers. These access factors likely increased individuals' opportunities to engage in online deviance. Additionally, those who hacked without detection were more likely to have engaged in piracy over the last year.

Those who hacked without detection were also more likely to have lower levels of self-control, lower parental supervision, and lower bonds to family. These conditions likely increased individuals' willingness to engage in wrongdoing and decreased their perceived risk of detection.

Individuals who hacked without detection were also more likely to be male and have a family car. Age group was approaching significance (0.065), with individuals in higher age groups demonstrating a greater likelihood of hacking. Living in a small town was not significant in the model, suggesting no geographic difference between the two groups.

The use of technology was not significantly different between those who hacked and were detected and those who did not hack. The only difference between these two groups with respect to opportunity variables were that they were more likely to spend time with peers and engage in piracy.

Additionally, those who were detected had lower levels of self-control and weaker family bonds compared to those who did not hack. The fact that parental supervision was non-significant, as were the technology use variables, suggests that those who hacked may have acted on opportunities to offend but were more likely to be observed compared to those who hacked without detection.

Lastly, individuals who were detected were more likely to be male and live in smaller towns. This relationship reflects both the observed gender differences in hacking, as well as potential differences in the likelihood of detection for individuals who reside in smaller geographic areas.

DISCUSSION AND CONCLUSION

Research examining juvenile delinquency highlights the need to deter future wrongdoing through detection and punishment of behavior (Nagin and Pogarsky, 2001; Pratt et al., 2006). The growth of the Internet and computer technology have created new platforms to engage in delinquent acts, many of which are difficult to observe in real time compared to traditional offline delinquency (Maimon et al., 2014; Marcum et al., 2014; Holt and Bossler, 2016). As a result, there is a need to consider the factors associated with the likelihood of detection for online offending among juveniles in order to develop better prevention and treatment programs (Holt and Bossler, 2016; NCA, 2017). This study attempted to address this issue through an examination of the behavioral and attitudinal correlates of juveniles' self-reported involvement in computer hacking and whether their behaviors were detected. A multinomial regression model was estimated using an international sample of juveniles collected through the ISRD-2 dataset (Junger-Tas and Marshall, 2012).

The findings demonstrated key support for all of the hypothesized relationships identified within the extant literature. First, low self-control was a significant predictor of hacking, regardless of whether the individual's behavior was detected. This finding is consistent with the broader hacking literature that show individuals with low self-control to be significantly more likely to engage in various hacking behaviors (Bossler and Burruss, 2011; Holt et al., 2012, 2019; Marcum et al., 2014; Udris, 2016). In fact, youth with low self-control were more likely to act on opportunities to hack, even in the face of detection from formal and informal sources of control as a result of their volatile temperament, impulsivity, self-centeredness, and risk-taking nature (Gottfredson and Hirschi, 1990; Bossler and Burruss, 2011).

This analysis also found partial support for opportunity factors and the risk of detection related to hacking. While having access to one's own computer and mobile phone were significantly related to hacking undetected, having a private bedroom was non-significant in both models. As a result, having one's own device may be a bigger factor in reducing the risk of detection compared to having a private physical space in which to operate (Jordan and Taylor, 1998; Holt, 2007; Steinmetz, 2016). If individuals must utilize a shared computer, it may increase the risk of detection due to the introduction of new programs or hardware and software that may be needed in order to hack. This is reinforced by the fact that there were no differences in technology ownership and use behaviors between those whose hacking behaviors were detected and those who did not hack. In much the same way, respondents who reported engaging in piracy were significantly more likely to hack, regardless of whether their activities were identified (Holt and Copes, 2010; Bossler and Burruss, 2011; Holt et al., 2012). Thus, greater access to and use of technology may decrease an individual's risk of detection for hacking generally.

In addition, time spent with peers was a significant predictor of hacking behavior, regardless of the likelihood of detection. The significant influence of delinquent peers on individual offending has been consistently identified in research on delinquency online (Bossler and Burruss, 2011; Holt et al., 2012; Marcum et al., 2014) and offline (Osgood et al., 1996; Haynie and Osgood, 2005; Hoeben et al., 2016). In fact, spending time with friends can provide models for offending and justifications for delinquency that increase an individual's risk of offending generally. This finding is compounded by the significant relationship identified between diminished parental supervision and undetected hacking. If parents do not know who their child spends time with, they may be more likely to socialize with delinquent peers (Hirschi, 1969; Sampson and Laub, 2003; Posick and Rocque, 2015).

The role of weakened family bonds and diminished supervision was also significantly associated with hacking without detection. This finding is consistent with previous research as those with weak parental attachments were at greater risk of engaging in deviance (Hirschi, 1969; Gottfredson and Hirschi, 1990; Wright et al., 1999; Sampson and Laub, 2003; Li, 2004; Posick and Rocque, 2015; Udris, 2016; Back et al., 2018). The role of parental bonds with respect to hacking is particularly salient as youth seem most likely to hack while at home due to ease of access to computers and greater uninterrupted time while using the device. The absence of significant differences between those who did not hack and those whose hacks were detected suggests the need for parental attachments and youth involvement in order to decrease the risk of juvenile hacking, similar to traditional delinquency.

The study also found several demographic factors associated with hacking. Those whose families owned a car were more likely to hack undetected, which may be a proxy for differential opportunities to use technology as a function of economic advantage (Schell and Dodge, 2002; Holt, 2007; Steinmetz, 2016; Holt et al., 2019). Males were also more likely to hack, whether detected or undetected, consistent with both previous quantitative and qualitative studies on hacking behaviors (Gilboa, 1996; Jordan and Taylor, 1998; Taylor, 1999; Schell and Dodge, 2002; Hutchings and Chua, 2017; Holt et al., 2019). It is unclear if this dynamic reflects differential supervision of behavior based on gender (Daigle et al., 2007; Lanctôt and Guay, 2014), or more unique factors associated with computer hacking generally. Lastly, youth living in smaller cities were more likely to have their hacking detected. This may be a function of reduced opportunities for unstructured socialization, as well as greater social bonds to parents as identified in prior research (Gardner and Shoemaker, 1989). These dynamics require further research in order to understand the role of demographic factors in the risk of online offending generally (Hutchings and Chua, 2017; Holt et al., 2019).

This study has direct implications for the development of programs to reduce juvenile hacking, as few have considered the factors that may increase the potential for obfuscation or detection of computer hacking (Holt and Bossler, 2016; NCA, 2017). The findings from the multinomial regression models demonstrated that hacking has some unique qualities that differentiate it from offline offending (see Bossler and Burruss, 2011; Steinmetz, 2016), but shared behavioral and attitudinal factors similar to that of traditional delinquency. As a result, there may be no need for specialized delinquency prevention programs for cybercrime. Instead, practitioners may benefit from incorporating information regarding simple forms of computer hacking into existing programmatic materials. Additionally, there is a need to increase parental awareness of cybercrime as a form of juvenile delinquency so as to improve the degree of supervision and oversight that may reduce opportunities to hack (Holt et al., 2012; Marcum et al., 2014). Lastly, substantive empirical research is needed to develop and evaluate the success of any prevention program that may emerge, whether in traditional delinquency reduction programs or those unique to cybercrime generally (Holt and Bossler, 2016; Leukfeldt, 2017; NCA, 2017).

Though this study provides an examination of an understudied issue associated with juvenile hacking, there are several limitations that must be noted. First, these data were collected between 2005 and 2007 when both the Internet and computer technology were less advanced and more costly. Future research would benefit from exploring whether the significant relationships identified in this analysis are also present in a more contemporary sample of youth. Relatedly, the current study is limited by its use of a predominately Western sample population. Future research should explore whether these factors are differentially associated with hacking and detection among Asian, Oceanic, African, and other nationally representative populations (Holt and Bossler, 2016).

The cross-sectional design of this study also presents some limitations as to the theoretical implications of this analysis. Cross-sectional studies provide important information regarding significant relationships between concepts and variables, though longitudinal data is needed to advance understanding of the temporal causes, pathways, and trends of juvenile hacking and detection (Holt et al., 2012; Marcum et al., 2014; Udris, 2016; NCA, 2017). The secondary nature of the data also limited the potential to examine the nature of the hacks reported by respondents, or their technical skills. It may be that individuals who engaged in more sophisticated or ethical hacks were able to continue without detection or sanction from formal and/or informal sources of social control. Furthermore, the dataset contained no measures regarding peer hacking behaviors, restricting the current study's operationalization of peer association (Bossler and Burruss, 2011; Holt et al., 2012; Marcum et al., 2014). Such information is essential in improving our understanding of the nature of hacking and its similarities to traditional offline delinquency.

DATA AVAILABILITY STATEMENT

The Second International Self-Report of Delinquency (ISRD-2) dataset is in the ICPSR repository.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

JL developed the primary idea, and wrote the introduction and literature review. TH conducted the statistics, wrote the methods and findings. Both wrote the discussion and conclusions and provided final revisions.

REFERENCES

Bachmann, M. (2010). The risk propensity and rationality of computer hackers. *Int. J. Cyber Criminol.* 4, 643–656.

Back, S., Soor, S., and LaPrade, J. (2018). Juvenile hackers: an empirical test of self-control theory and social bonding theory. *Int. J. Cybersecur. Intell. Cybercrime* 1, 40–55.

Bae, S. M. (2017). The influence of strain factors, social control factors, self-control and computer use on adolescent cyber delinquency: Korean National Panel Study. *Child. Youth Serv. Rev.* 78, 74–80. doi: 10.1016/j.childyouth.2017.05.008

Bossler, A. M., and Burruss, G. W. (2011). "The general theory of crime and computer hacking:Low self-control hackers?," in *Corporate Hacking and Technology-Driven Crime: Social Dynamics and Implica-tions*, eds T. J. Holt and B. H. Schell (Hershey, PA: IGI Global), 38–67. doi: 10.4018/978-1-61692-805-6.ch003

Botchkovar, E., Marshall, I. H., Rocque, M., and Posick, C. (2015). The importance of parenting in the development of self-control in boys and girls: Results from a multinational study of youth. *J. Crim. Justice* 43, 133–141. doi: 10.1016/j.jcrimjus.2015.02.001

Calce, M., and Silverman, C. (2008). *Mafiaboy: How I Cracked the Internet and why It's Still Broken*. Toronto: Penguin Group.

Cardone, C., and Hayes, R. (2012). Shoplifter perceptions of store environments: an analysis of how physical cues in the retail interior shape shoplifter behavior. *J. Appl. Secur. Res.* 7, 22–58. doi: 10.1080/19361610.2012.631178

Cherbonneau, M., and Copes, H. (2005). 'Drive it like you stole it' auto theft and the illusion of normalcy. *Br. J. Criminol.* 46, 193–211. doi: 10.1093/bjc/azi059

Clarke, R. V. G. (ed.) (1997). *Situational Crime Prevention*. Monsey, NY: Criminal Justice Press, 225–256.

Cohen, L. E., and Felson, M. (1979). Social change and crime rate trends: a routine activity approach. *Am. Sociol. Rev.* 44, 588–608.

Cohen, P., Cohen, J., Aiken, L. S., and West, S. G. (1999). The problem of units and the circumstance for POMP. *Multivariate Behav. Res.* 34, 315–346. doi: 10.1207/s15327906mbr3403_2

Coleman, E. G. (2014). *Hacker, Hoaxer, Whistleblower, Spy: The Many Faces of Anonymous*. Brooklyn, NY: Verso Books.

Cornish, D. B., and Clarke, R. V. (2003). Opportunities, precipitators and criminal decisions: a reply to Wortley's critique of situational crime prevention. *Crime Prev. Stud.* 16, 41–96.

Cornish, D. B., and Clarke, R. V. (2014). *The Reasoning Criminal: Rational Choice Perspectives on Offending*. London: Transaction Publishers.

Daigle, L. E., Cullen, F. T., and Wright, J. P. (2007). Gender differences in the predictors of juvenile delinquency: assessing the generality-specificity debate. *Youth Violence Juv. Justice* 5, 254–286. doi: 10.1177/1541204007301289

Felson, M. (1986). "Routine activities, social controls, rational decisions and criminal outcomes," in *The Reasoning Criminal*, eds D. Cornish and R. V. Clarke (New York, NY: Springer-Verlag), 302–327.

Felson, M. (1995). Those who discourage crime. *Crime Place* 4, 53–66.

Furnell, S. (2002). *Cybercrime: Vandalizing the Information Society*. London: Addison-Wesley, 3–540.

Gardner, L., and Shoemaker, D. J. (1989). Social bonding and delinquency: a comparative analysis. *Sociol. Q.* 30, 481–499. doi: 10.1111/j.1533-8525.1989.tb01532.x

Gilboa, N. (1996). "Elites, lamers, narcs, and whores: exploring the computer under-ground," in *Wired Women: Gender and New Realities in Cyberspace*, eds L. Cherny and E. Weise (Seattle, WA: Seal Press), 98–113.

Gottfredson, M. R., and Hirschi, T. (1990). *A General Theory of Crime*. Palo Alto, CA: Stanford University Press.

Grabosky, P. (2016). "The evolution of cybercrime, 2006–2016," in *Cybercrime through Aninterdisciplinary Lens*, ed. T. J. Holt (New York, NY: Routledge), 29–50.

Grasmick, H. G., Tittle, C. R., Bursik, R. J. Jr., and Arneklev, B. J. (1993). Testing the coreempirical implications of Gottfredson and Hirschi's general theory of crime. *J. Res. Crime Delinq.* 30, 5–29.

Haynie, D. L., and Osgood, D. W. (2005). Reconsidering peers and delinquency: How do peers matter? *Soc. Forces* 84, 1109–1130. doi: 10.1353/sof.2006.0018

Hirschi, T. (1969). A control theory of delinquency. *Criminol. Theory* 1969, 289–305.

Hoeben, E. M., Meldrum, R. C., Walker, D., and Young, J. T. (2016). The role of peer delinquency and unstructured socializing in explaining delinquency and sub-stance use: a state-of-the-art review. *J. Crim. Justice* 47, 108–122. doi: 10.1016/j.jcrimjus.2016.08.001

Hollinger, R. C., and Lanza-Kaduce, L. (1988). The process of criminalization: The case of computer crime laws. *Criminology* 26, 101–126. doi: 10.1111/j.1745-9125.1988.tb00834.x

Holt, T. J. (2007). Subcultural evolution? Examining the influence of on-and off-line experiences on deviant subcultures. *Deviant Behav.* 28, 171–198. doi: 10.1080/01639620601131065

Holt, T. J. (2010). Examining the role of technology in the formation of deviant subcultures. *Soc. Sci. Comput. Rev.* 28, 466–481. doi: 10.1177/0894439309351344

Holt, T. J., and Bossler, A. M. (2016). *Cybercrime In Progress: Theory and Prevention of Technology-Enabled Offenses*. London: Routledge.

Holt, T. J., Bossler, A. M., and May, D. C. (2012). Low self-control, deviant peer associations, and juvenile cyberdeviance. *Am. J. Crim. Justice* 37, 378–395. doi: 10.1007/s12103-011-9117-3

Holt, T. J., Burruss, G. W., and Bossler, A. M. (2010). Social learning and cyber-deviance: examining the importance of a full social learning model in the virtual world. *J. Crime Justice* 33, 31–61. doi: 10.1080/0735648x.2010.9721287

Holt, T. J., and Copes, H. (2010). Transferring subcultural knowledge on-line: Practices and beliefs of persistent digital pirates. *Deviant Behav.* 31, 625–654. doi: 10.1080/01639620903231548

Holt, T. J., Navarro, J. N., and Clevenger, S. (2019). Exploring the moderating role of gender in juvenile hacking behaviors. *Crime Delinq.* doi: 10.1177/0011128719875697

Hutchings, A., and Chua, Y. T. (2017). "Gendering cybercrime," in *Cybercrime through an Interdisciplinary Lens*, ed. T. J. Holt (Oxon: Routledge), 181–202.

Jordan, T., and Taylor, P. (1998). A sociology of hackers. *Sociol. Rev.* 46, 757–780. doi: 10.1111/1467-954x.00139

Junger-Tas, J. (2010). The significance of the international self-report delinquency study (ISRD). *Eur. J. Crim. Policy Res.* 16, 71–87. doi: 10.1007/s10610-010-9119-6

Junger-Tas, J., and Marshall, I. H. (2012). "Introduction to the international self-report study of delinquency (ISRD-2)," in *The Many Faces of Youth Crime*, eds J. Junger-Tas I, H. Marshall, D. Enzmann, M. Killias, M. Steketee, and

B. Gruszczynska (New York, NY: Springer), 3–20. doi: 10.1007/978-1-4419-9455-4_1

Kong, J., and Lim, J. (2012). The longitudinal influence of parent–child relationships and depression on cyber delinquency in South Korean adolescents: a latent growth curve model. *Child. Youth Serv. Rev.* 34, 908–913. doi: 10.1016/j.childyouth.2012.01.020

Lanctôt, N., and Guay, S. (2014). The aftermath of workplace violence among health-care workers: a systematic literature review of the consequences. *Aggress. Violent Behav.* 19, 492–501. doi: 10.1016/j.avb.2014.07.010

Leukfeldt, E. R. (Ed.). (2017). *The Human Factor in Cybercrime and Cybersecurity: Research Agenda*. The Netherlands: Eleven International Publishing.

Levy, S. (1984). *Hackers: Heroes of the Computer Revolution*, Vol. 14. Garden City, NY: Anchor Press.

Li, S. D. (2004). The impacts of self-control and social bonds on juvenile delinquency in a national sample of midadolescents. *Deviant Behav.* 25, 351–373. doi: 10.1080/01639620490441236

Maimon, D., Alper, M., Sobesto, B., and Cukier, M. (2014). Restrictive deterrent effects of a warning banner in an attacked computer system. *Criminology* 52, 33–59. doi: 10.1111/1745-9125.12028

Marcum, C. D., Higgins, G. E., Ricketts, M. L., and Wolfe, S. E. (2014). Hacking in high school: cybercrime perpetration by juveniles. *Deviant Behav.* 35, 581–591. doi: 10.1080/01639625.2013.867721

Marshall, I. H., and Enzmann, D. (2012). "Methodology and design of the ISRD-2 study," in *The Many Faces of Youth Crime*, eds J. Junger-Tas I, H. Marshall, D. Enzmann, M. Killias, M. Steketee, and B. Gruszczynska (New York, NY: Springer), 21–65. doi: 10.1007/978-1-4419-9455-4_2

Nagin, D. S., and Pogarsky, G. (2001). Integrating celerity, impulsivity, and extralegal sanction threats into a model of general deterrence: Theory and evidence. *Criminology* 39, 865–892. doi: 10.1111/j.1745-9125.2001.tb00943.x

NCA (2017). *Intelligence Assessment: Pathways into Cybercrime*. London: National Crime Agency.

Osgood, D. W., Wilson, J. K., O'Malley, P. M., Bachman, J. G., and Johnston, L. D. (1996). Routine activities and individual deviant behavior. *Am. Sociol. Rev.* 61, 635–655.

Posick, C., and Rocque, M. (2015). Family matters: a cross-national examination of family bonding and victimization. *Eur. J. Criminol.* 12, 51–69. doi: 10.1177/1477370814538777

Pratt, T. C., and Cullen, F. T. (2000). The empirical status of Gottfredson and Hirschi's general theory of crime: a meta-analysis. *Criminology* 38, 931–964. doi: 10.1111/j.1745-9125.2000.tb00911.x

Pratt, T. C., Cullen, F. T., Blevins, K. R., Daigle, L. E., and Madensen, T. D. (2006). "The empirical status of deterrence theory: a meta-analysis," in *Taking Stock: The Status of Criminological Theory Advances in Criminological Theory*, 15th Edn, eds F. T. Cullen, J. P. Wright, and K. R. Blevins (New Brunswick, NJ: Transaction Publishing), 367–396.

Reyns, B. W. (2010). A situational crime prevention approach to cyberstalking victimization: preventive tactics for Internet users and online place managers. *Crime Prev. Community Saf.* 12, 99–118. doi: 10.1057/cpcs.2009.22

Rocque, M., Posick, C., and Zimmerman, G. M. (2013). Measuring up: Assessing themeasurement properties of two self-control scales. *Deviant Behav.* 34, 534–556. doi: 10.1080/01639625.2012.748619

Rogers, M., Smoak, N. D., and Liu, J. (2006). Self-reported deviant computer behavior: a big-5,moral choice, and manipulative exploitive behavior analysis. *Deviant Behav.* 27, 245–268. doi: 10.1080/01639620600605333

Sampson, R. J., and Laub, J. H. (2003). Life-course desisters? Trajectories of crime amongdelinquent boys followed to age 70. *Criminology* 41, 555–592. doi: 10.1111/j.1745-9125.2003.tb00997.x

Schell, B. H., and Dodge, J. L. (2002). *The Hacking of America: Who's Doing it, Why, and How*. Westport, CT: Greenwood Publishing Group Inc.

Skinner, W. F., and Fream, A. M. (1997). A social learning theory analysis of computer crimeamong college students. *J. Res. Crime Delinq.* 34, 495–518.

Slatalla, M., and Quittner, J. (1995). *Masters of Deception*. New York, NY: Harper Collins.

Steinmetz, K. F. (2015). Craft (y)ness. An ethnographic study of hacking. *Br. J. Criminol.* 55, 125–145. doi: 10.1093/bjc/azu061

Steinmetz, K. F. (2016). *Hacked: A Radical Approach to Hacker Culture and Crime*. New York, NY: NYU Press.

Steinmetz, K. F. (2017). Ruminations on warning banners, deterrence, and system intrusionresearch. *Criminol. Public Policy* 16, 727–737. doi: 10.1111/1745-9133.12314

Taylor, P. (1999). *Hackers: Crime and the Digital Sublime*. New York, NY: Routledge.

Udris, R. (2016). Cyber deviance among adolescents and the role of family, school, and neighborhood: a cross-national study. *Int. J. Cyber Criminol.* 10, 127–146.

Vazsonyi, A. T., Mikuška, J., and Kelley, E. L. (2017). It's time: a meta-analysis on the self-control deviance link. *J. Crim. Justice* 48, 48–63. doi: 10.1016/j.jcrimjus.2016.10.001

Wall, D. S. (2001). "Cybercrimes and the internet," in *Crime and the Internet*, ed. D. S. Wall (New York, NY: Routledge), 1–17. doi: 10.4324/9780203164501_chapter_1

Wilson, T., Maimon, D., Sobesto, B., and Cukier, M. (2015). The effect of a surveillance banner in an attacked computer system: additional evidence for the relevance of restrictive deterrence in cyberspace. *J. Res. Crime Delinq.* 52, 829–855. doi: 10.1177/0022427815587761

Wright, B. R. E., Caspi, A., Moffitt, T. E., and Silva, P. A. (1999). Low self-control, social bonds, and crime: Social causation, selection, or both? *Criminology* 37, 479–514. doi: 10.1111/j.1745-9125.1999.tb00494.x

Wright, R. T., and Decker, S. H. (1996). *Burglars on the Job: Streetlife and Residential Break-ins*. Lebanon, NH: UPNE.

Yar, M. (2013). *Cybercrime and Society*. London: SAGE Publications.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ORIGINAL RESEARCH
published: 16 June 2020
doi: 10.3389/fpsyg.2020.01049

Cognitive Models in Cybersecurity: Learning From Expert Analysts and Predicting Attacker Behavior

Vladislav D. Veksler[1]*, *Norbou Buchler*[2], *Claire G. LaFleur*[2], *Michael S. Yu*[3], *Christian Lebiere*[3] *and Cleotilde Gonzalez*[3]

[1] DCS Corporation, U.S. Army Data & Analysis Center, Aberdeen Proving Ground, MD, United States, [2] U.S. Army Data & Analysis Center, Aberdeen Proving Ground, MD, United States, [3] Department of Psychology, Carnegie Mellon University, Pittsburgh, PA, United States

Cybersecurity stands to benefit greatly from models able to generate predictions of attacker and defender behavior. On the defender side, there is promising research suggesting that Symbolic Deep Learning (SDL) may be employed to automatically construct cognitive models of expert behavior based on small samples of expert decisions. Such models could then be employed to provide decision support for non-expert users in the form of explainable expert-based suggestions. On the attacker side, there is promising research suggesting that model-tracing with dynamic parameter fitting may be used to automatically construct models during live attack scenarios, and to predict individual attacker preferences. Predicted attacker preferences could then be exploited for mitigating risk of successful attacks. In this paper we examine how these two cognitive modeling approaches may be useful for cybersecurity professionals via two human experiments. In the first experiment participants play the role of cyber analysts performing a task based on Intrusion Detection System alert elevation. Experiment results and analysis reveal that SDL can help to reduce missed threats by 25%. In the second experiment participants play the role of attackers picking among four attack strategies. Experiment results and analysis reveal that model-tracing with dynamic parameter fitting can be used to predict (and exploit) most attackers' preferences 40 – 70% of the time. We conclude that studies and models of human cognition are highly valuable for advancing cybersecurity.

Keywords: cyber-security, cognitive modeling, behavioral simulations, deep learning, reinforcement learning, decision support, XAI (eXplainable Artificial Intelligence), human-agent teaming

OPEN ACCESS

Edited by:
Paul Watters,
La Trobe University, Australia

Reviewed by:
Char Sample,
Idaho National Laboratory (DOE),
United States
Stefan Sütterlin,
Østfold University College, Norway

****Correspondence:***
Vladislav D. Veksler
vdv718@gmail.com

Specialty section:
This article was submitted to
Cognition,
a section of the journal
Frontiers in Psychology

Received: *17 January 2020*
Accepted: *27 April 2020*
Published: *16 June 2020*

Citation:
Veksler VD, Buchler N, LaFleur CG, Yu MS, Lebiere C and Gonzalez C (2020) Cognitive Models in Cybersecurity: Learning From Expert Analysts and Predicting Attacker Behavior. Front. Psychol. 11:1049. doi: 10.3389/fpsyg.2020.01049

1. INTRODUCTION

The field of cybersecurity has as much to do with human agency as it does with computer network integrity. However, while computer network technology changes rapidly on a regular basis, human learning and decision mechanisms do not. This being the case, research focused on cognitive science may provide the needed breakthrough capabilities for long-term network security and a greater return on investment than efforts chasing the latest software vulnerabilities. Cognitive science, and cognitive modeling in particular show a great promise for the field of cybersecurity (Veksler et al., 2018).

The goal of this paper is to provide examples of how computational models of human cognition may be employed to predict human preferences and behavior in cybersecurity. This paper focuses on two specific examples of the use of cognitive modeling in the context of cybersecurity. On the defender side, we aim to construct cognitive models of cyber analysts working with Intrusion Detection Systems (IDS) based on expert behavior, and employ these models to provide suggestions to non-expert analysts. On the attacker side, we aim to construct models of individual attacker decision biases, and employ these models to reduce the risk of successful attacks.

IDS software provides analysts with aggregated logs of network-activity alert records, where each record includes a number of threat-relevant features, and the job of a cyber analyst is to either elevate an alert as a potential threat, or to dismiss it as a false alarm. Predicting expert cyber analyst behavior in such a domain presents some challenges (Gonzalez et al., 2014). Traditional approaches in Computer Science routinely employ some Machine Learning (ML) classifier, training it on expert decisions with alert record features being classifier inputs, and the threat/no-threat classification being the output. Deep Learning (DL) methodology in particular has gained much acclaim in the recent decade as a successful technology for classifying large noisy complex data. The problem is that the availability of labeled cyber expert decision data is fairly sparse, often comprising just a few dozen or hundreds of examples (whereas DL requires training data with thousands or millions of examples). Additionally, DL-based recommendations are not easily explainable, and thus may not be well-suited for decision-aid software. Symbolic Deep Learning (SDL) may be a better approach for constructing models of expert behavior (Veksler and Buchler, 2019). The advantage of this approach is that it addresses the challenge of developing flexible and explainable models of cognition and behavior based on small samples of data.

Attacker-defender dynamics are often modeled in terms of Game Theory (GT). Game-theoretic approaches are useful for determining optimal mixes of strategies for leaving an attacker without a preferred strategy of their own. Moreover, GT-based defense algorithms have been successfully applied in many real-world security scenarios, including airport security, coast guard, police, and anti-poaching (animal preservation) efforts (Tambe et al., 2014). Veksler and Buchler (2016) and Cranford et al. (2019) argue that cognitive modeling techniques, and more specifically model-tracing[1] and dynamic parameter fitting, may be used to track individual attacker preferences in real time, providing fairly high improvements over normative GT approaches in reducing the potential for successful attacks.

The rest of this paper presents two experiments with respective simulations and analyses, specifically aimed at examining the two uses of cognitive modeling in cybersecurity described above. In the first experiment participants play the role of cyber analysts performing a task based on IDS alert elevation. The SDL-based cognitive models are trained on data from top-performing participants. Results from all other participants are examined for degree of potential reduction in missed threats based on the trained SDL models. In the second experiment participants play the role of attackers picking among four attack strategies, and playing against defenders based on either normative game theory or against adaptive cognitive models using model-tracing with dynamic parameter fitting. Results from this experiment are analyzed for the degree to which human attacker preferences can be predicted and exploited.

2. CONSTRUCTING COGNITIVE MODELS OF EXPERT ANALYSTS

A large subset of cybersecurity professionals are analysts working with Intrusion Detection Systems (IDS). This is often a grueling job requiring constant real-time monitoring of incoming network alerts for 12 hours straight (panama shifts). Employee turnover in these jobs is very high, in part because panama shifts are often incompatible with human health, mental health, and family life (Stimpfel et al., 2012; Oltsik, 2019). The job requires each analyst to sift through incoming alerts, picking out which alerts are worth elevating, and which are false alarms (D'Amico and Whitley, 2008). There is no way to train someone for this job via standard schooling, because every network has its own particularities, so all training is on-the-job training, even for experienced IDS professionals. In other words, this is a field where employees take a long time to gain expertise, and leave fairly soon after gaining said expertise.

Just like in the medical industry, the largest number of errors for IDS analysts happens after shift changes (which is the reason why panama shifts are the industry standard—to minimize the number of shift changes) (Friesen et al., 2008). After an employee spends 12 hours gaining expertise for the context of that day's alerts, they go home and leave the job to someone who is now missing all of the context needed for correct alert identification. The prescribed operating procedure is for each analyst to leave notes for the next shift, and to read notes left by analysts working the prior shift. However, based on our interviews with cyber-analysts, we found out that this rarely happens. Some analysts are better at taking and leaving notes than others, but there is always information lost between shifts.

A useful tool one might design for cyber-analysts would be a decision-aid that could highlight the alerts that an expert might elevate (whether we are talking about long-term expertise of veteran professionals, or more localized expertise relating to the state of recent network activity). This may be useful to enable new employees to see potential decisions of veteran employees, or it may be useful to enable analysts starting a shift to see potential decisions of analysts from the prior shift. In either case what is required is to train a model on contextualized expert decision-making and use it to predict future alert elevation.

In Machine Learning terms, the problem may be framed as a supervised classification problem where a model is trained on expert behavior, and its predictions are used for suggestions. Deep Learning in particular has garnered much attention over the last decade as being a highly successful ML technique for classification of complex and noisy data (e.g., Rusk, 2015). However, DL requires much larger training sets than are available in the context of expert analysts. Moreover, DL approaches are largely unexplainable and prone to catastrophic interference.

[1] Model-tracing comprises force-feeding individual human experiences into the model. More on this in Dynamic Cognitive Models of Attacker Preferences section.

That is, (1) there is no way for an analyst to ask "why" a specific alert was highlighted, and (2) updating the model with new expert decision data could cause the model to "forget" prior learned classifications.

Veksler and Buchler (2019) propose Symbolic Deep Learning (SDL) as an alternative approach to constructing user models. A symbolic version of deep learning is promising in that this method is capable of building classifiers from a small number of examples (Dutra et al., 2017; Zhang and Sornette, 2017; d'Avila Garcez et al., 2018). In this way, SDL learning efficiency is more akin to that of humans, and SDL is much more appropriate for creating models from individual or small-group data than DL. Whereas, DL builds up a black-box model of behavior, SDL builds up an explainable model of expert cognition—an expandable hierarchical memory network based on expert experiences and decisions.

More specifically, a traditional deep neural network has a pre-specified number of layers, with a pre-specified number of nodes in each layer, and with nodes of each pair of successive layers being fully interconnected via weighted links (see **Figure 1**, left). As the network learns, the weights on these links change, but at no point can one look at those links and comprehend what exactly the network has learned. Because all knowledge is distributed among the links, the network has to be large enough to be able to learn a given problem, and thus requires thousands or millions of iterations to learn even simple input-output mapping. Symbolic deep nets, on the other hand, start with no nodes between input and output layers, and learn these nodes based on perceived input node co-occurrences (see **Figure 1**, middle and right). These deep nodes are essentially combinations of input features (a.k.a., chunks or configural cues), and in the case of modeling human memory, deeper chunks are taken to represent deeper domain expertise (Feigenbaum and Simon, 1984; Gobet, 1998). Due to the symbolic nature of chunk-based hierarchical memory, one can look at the learned chunks at any time so as to gain insight into what the network has learned. Because of the nature of chunk learning (one-shot learning), simple feature combinations can be learned quickly, enabling symbolic nets to learn at speeds on par with human learning—from just a few examples, rather than tens of thousands.

The major hurdle for symbolic deep models of memory has been a combinatoric explosion of memory. For example, the configural-cue model of memory (Gluck and Bower, 1988) creates a configural node (i.e., chunk) for every unique set of potential inputs, thus creating a maximum of $(k+1)^n-1$ memory chunks, where n is the number of input dimensions and k is the number of possible input values along each input dimension[2]. However, this problem is alleviated when chunks are created in a more conservative manner. For example, Veksler et al. (2014) employed the ACT-R (Anderson, 1993; Anderson and Lebiere, 1998) rational memory activation mechanism, where memory activation is based on its recency/frequency of use, as a selection mechanism for which memory nodes could be chunked.

In Experiment 1 below we gather data from participants classifying threats in an IDS-like environment, so as to examine how SDL-based cognitive models and DL-based behavior models may be able to learn from smallish data sets of expert behavior in such an environment, and to what degree these methods may be helpful in highlighting alerts for non-expert participants. Specifically, for the simulations below we employ a popular DL framework TensorFlow (Abadi et al., 2016) and the conservative-rational SDL framework (Veksler et al., 2014). The conservative-rational framework was originally proposed as an amalgamation of two other models of symbolic hierarchical memory—the configural-cue memory structure (Gluck and Bower, 1988), and the ACT-R cognitive architecture chunk activation mechanism (Anderson, 1993; Anderson and Lebiere, 1998)—for the purposes of combining the category-learning abilities of the configural-cue model and the computational efficiency of rational memory activation in ACT-R.

The purpose of the simulations below is to provide evidence that the SDL cognitive modeling technique may be useful in the context of aiding security analysts, rather than to find optimal model performance. Thus, we did not perform any parameter search for SDL, and merely used default framework parameters. For DL we attempted simulations with a few different network shapes, so as to establish a more fair comparison of DL and SDL, because network shape is of a very high importance to DL modeling (not so for SDL, since its shape changes automatically). We found that a five-layer network of the shape $\{input, 50, 100, 100, output\}$ performed better than networks of smaller or greater depth and networks of smaller or greater widths[3].

We also attempted different numbers of training epochs for both DL and SDL. Employing multiple training epochs is of high importance for DL when working with smaller datasets. Essentially, if you have 1,000 training samples, training the network for 100 epochs enables you to simulate a dataset of 100,000 samples. Using an overly high number of epochs comes at a cost of overfitting. That is, the model might begin to perform very well on the training data, but will fail to generalize to a new dataset (test data) if the number of epochs is overly high. This is not as important for SDL, as it requires less than ten epochs to reach peak performance even on small datasets, but it is important nonetheless. All simulation results reported below are based on best-performing numbers of epochs for each model.

All simulation results below are averaged over one hundred simulation runs.

[2] Given n features (e.g., *large*, *square*, *white*), we can create a chunk for every combination of feature presence and absence ($\{large\}$, $\{square\}$, $\{white\}$, $\{large, square\}$, $\{large, white\}$, $\{square, white\}$, and $\{large, square, white\}$). If we represent feature presence as a 1 and feature absence as a 0, we can represent each chunk as a binary number, and the total number of possible chunks is the total number of possible binary numbers, minus the blank chunk, which is 2^n-1. When each feature dimension can have two potential values, the total number of possible chunks is $3^n - 1$. With k possible values on n feature dimensions, we can have at most $(k + 1)^n - 1$ possible chunks to represent all potential feature combinations.

[3] All performance comparisons were in terms of the sensitivity heuristic, d'. Network layer depths were incremented and decremented by one layer for comparisons. Network layer widths were incremented and decremented by 25–50 nodes for comparisons.

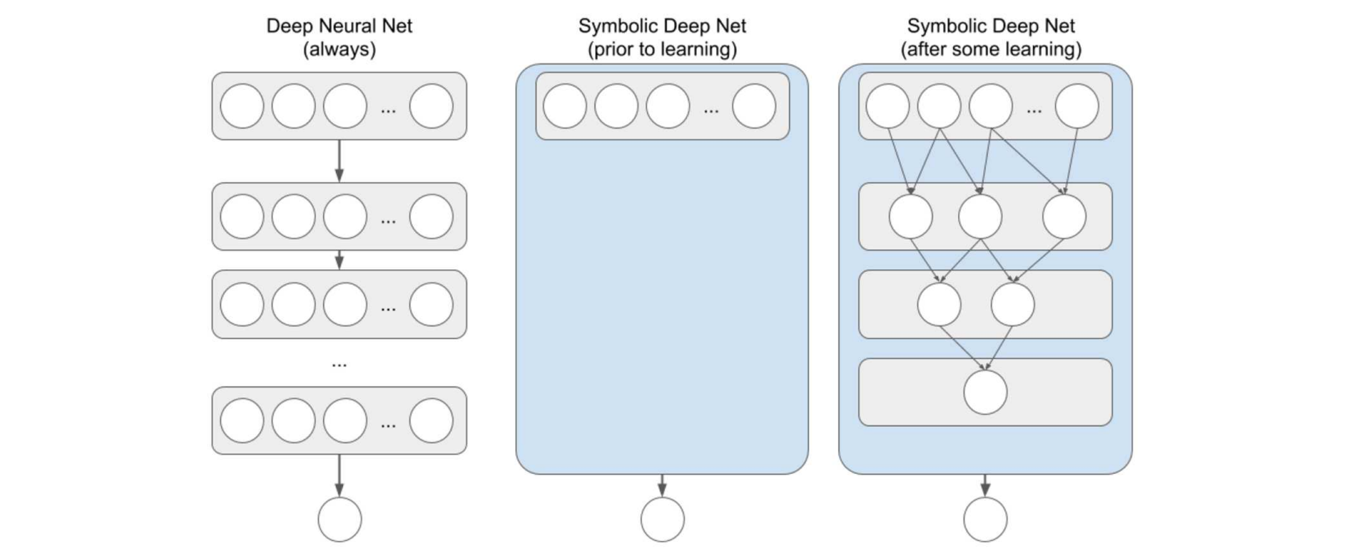


FIGURE 1 | Traditional Deep Neural Net and Symbolic Deep Net structures for networks with a single output node. Each circle represents a network node. Top row of each net represents the input layer, bottom row is output nodes. Thicker arrows going from a node container to a node or another container represent a fully interconnected vector/matrix of weighted feed-forward links. Thinner arrows between nodes represent symbolic (i.e., not weighted) feed-forward links.

TABLE 1 | Experiment 1 threat classification rules.

Events are threats if they meet all four criteria:
1. Time Stamp between 0:00 and 5:00 h
2. Source Port < 80 or > 5,000
3. Destination Countries: Russia, China
4. Alerts including suspicious, encrypted, exploit, and virus

2.1. Experiment 1

Experiment 1 was designed as a part of a larger study (unpublished) to examine human ability to evaluate cyber-threats in a simplified IDS-like environment based on a small set of instructions. In this experiment participants were presented with four batches of cyber activity records, where Batch 1 had 40 records, Batch 2 had 60 records, Batch 3 had 80 records, and Batch 4 had 100 records. Batch order was randomized. Each record consisted of four features relating to the detected network activity: time stamp, source port number, country, and alert description (e.g., "FTP - Suspicious MGET Command," "ET TROJAN Qhosts Trojan Check-in").

For each such record participants were able to click either "Threat" or "No Threat" radio button. Half of the records in each batch were threats. Participants were not provided feedback as to whether their threat classifications were correct, however, threat classification rules (see **Table 1**) were always visible to the participants.

This study was performed online, using Amazon Mechanical Turk to recruit adult residents of the United States for pay. We recruited sixty one participants for this experiment.

2.1.1. Results

For the purposes of all model analyses below we employ Batches 1, 2, and 3 as model learning data (i.e., training sets), and Batch 4 (containing 100 record cases) for examination (i.e., test set). The highest overall identification score for Batches 1, 2, and 3 was 0.883, and this score was achieved by four participants. We classify these four participants as "experts" and train SDL and DL models only on those participants' decisions from Batches 1, 2, and 3 (not Batch 4).

Average overall score on Batch 4 for non-expert participants was 71.3% (random-level behavior is 50%), with average hitrate (HR; correctly identified threats) being 0.725 and false-alarm rate (FA) being 0.298[4]. According to Signal Detection Theory (Swets, 1964, 1996), given the same training time we could have pushed participants toward a higher hitrate at the cost of increasing false-alarms or a lower false-alarm rate at the cost of a lower hitrate, depending on the perceived subjective utilities of hits, misses, false alarms, and correct rejections, though what would remain constant is their ability to discriminate what constitutes a threat – the sensitivity characteristic, d'. In this experiment the non-expert sensitivity on Batch 4 was $d' = 1.128$ (higher d' suggests higher sensitivity; d' for random-level behavior is 0).

SDL model trained[5] on expert decisions from Batches 1, 2, and 3 (180 cases × 4 experts) produced an average score of 86.4% on Batch 4, with average hitrate being 0.796 and average false-alarm rate being 0.069, $d' = 2.31$. DL model trained[6] on expert decisions from Batches 1, 2, and 3 produced an average score of 77.6% on Batch 4, with average hitrate being 0.701 and average false-alarm rate being 0.149, $d' = 1.57$. When trained on just

[4] All analyses include only Batch 4 cases where participants made a threat/no-threat classification. Not all presented cases were answered by all participants. Forty nine participants provided all 100 answers, four of the participants provided only 99 answers, one participants provided only 77 answers, and three other participants provided <20 answers each.

[5] 3 epochs.

[6] 100 epochs.

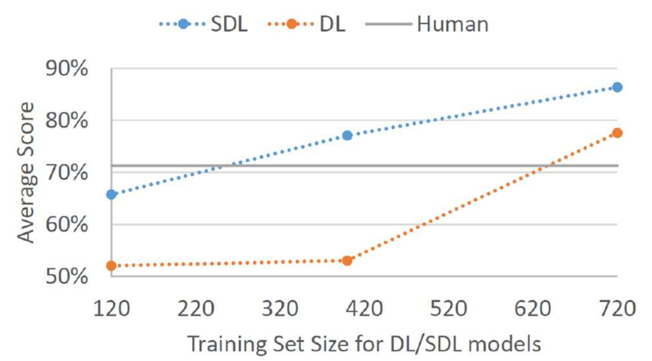


FIGURE 2 | Average correct classification score for SDL and DL models on Batch 4. Models were trained on Batch 1 of expert data (4 experts × 40 decisions each = 120 total cases), Batches 1 and 2 of expert data (4 experts × 100 decisions each = 400 total cases), and Batches 1, 2, and 3 of expert data (4 experts × 180 decisions each = 720 total cases). Gray baseline labeled "Human" represents average performance on Batch 4 for human non-expert participants.

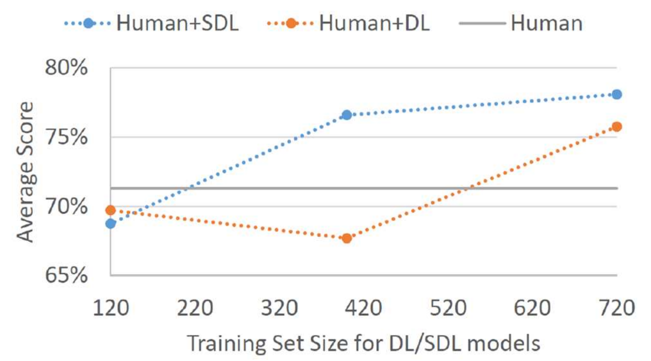


FIGURE 3 | Average correct classification score for non-experts on Batch 4 with the assumption that the non-experts would adopt all threat suggestions provided by a given helper-agent. The displayed results are for SDL and DL helper-agents that were trained on Batch 1 of expert data (4 experts × 40 decisions each = 120 total cases), Batches 1 and 2 of expert data (4 experts × 100 decisions each = 400 total cases), and Batches 1, 2, and 3 of expert data (4 experts × 180 decisions each = 720 total cases). Gray baseline labeled "Human" represents average performance on Batch 4 for non-expert participants without any helper-agent suggestions.

Batches 1 and 2 (100 cases × 4 experts) SDL[7] and DL[8] d' statistics fell to 1.76 and 0.16, respectively. When trained on just Batch 1 (40 cases × 4 experts) SDL[9] and DL[10] d' statistics fell to 0.84 and 0.14, respectively. **Figure 2** displays the overall performance scores for SDL and DL given the three training set sizes.

Perhaps more important than a standalone model score on Batch 4 is the degree to which such models can aid non-experts in their decision-making. Assuming that we employ SDL trained on expert decisions from Batches 1, 2, and 3 to highlight potential threats on Batch 4, and assuming that non-expert participants always classified the highlighted records as threats, non-expert hitrate would go up to 90.4%. This is a 25% improvement on correctly identified threats. This would come at a cost of false alarms rate increasing to 34.3% (15% increase); however, the overall ability to discriminate signal from noise for such human-agent teams would go from $d' = 1.13$ to $d' = 1.71$. Even if SDL was trained only on 100 decisions from each of the four experts (just Batches 1 and 2), the overall sensitivity to the threat signal would improve, $d' = 1.54$ ($HR = .870$, $FA = 0.338$). Employing SDL trained only on Batch 1 (40 decision from each of the four experts) would provide no decision improvement, $d' = 1.12$ ($HR = .861$, $FA = 0.486$). As would be expected based on standalone model performances, an analogous DL-based decision could help humans improve to a slightly lesser degree when trained on expert decisions from Batches 1, 2, and 3, $d' = 1.62$ ($HR = 0.913$, $FA = 0.397$), and not at all when trained only on Batches 1 and 2, $d' = 0.98$ ($HR = 0.801$, $FA = 0.447$), or only on Batch 1, $d' = 1.05$ ($HR = 0.767$, $FA = 0.372$). **Figure 3** displays the overall performance scores that can be achieved for non-expert human-agent teams given different sizes of expert decision training sets.

2.2. Discussion

One of the most clear results above is that SDL performs much better with smaller datasets than traditional DL methods. This matters a great deal in the field of cybersecurity, where expert data is difficult to come by, or where expertise is localized to a single 12-hour shift. Perhaps it should not be surprising that cognitive modeling methodology is more appropriate for building decision aids based on small samples of individual decisions than AI methods designed for large data mathematical optimization. Unfortunately, due to the popularity of traditional ML methodology, cognitive computing is often not taken into account, even when it may be the right tool for the job.

Note that the expert decisions that models were trained on were only 88.3% correct and SDL was able to achieve nearly this same performance, 86.4%, on the 100 test cases. If non-experts were to have complete trust that records highlighted as threats by the SDL decision aid are correct, their performance could improve from 71.3 to 78.1%. We could presume that if expert performance was better, and if more expert data was available, SDL performance and the degree to which it could help non-experts would improve, as well. On the flip side, if expert performance was worse, or if less expert data were available, we would expect worse performance from both SDL and DL. This suggests that experiments of this ilk may not replicate with smaller samples of participants. More importantly, selection of expert performers in the real world is of the highest concern for generating similar decision-aid training data.

One question that may come to mind is whether humans are needed at all. If it is the case that SDL performance is 86.4% whereas non-expert human performance is expected to be between 71.3 and 78.1% even with the SDL-based decision-aid, why not just train agents on expert behavior and let them loose without non-expert interference? However, this questions

[7] 6 epochs.
[8] 200 epochs.
[9] 8 epochs.
[10] 200 epochs.

presumes static non-expert performance, whereas humans learn and adapt. Human novices may begin with lower levels of performance, but when provided with expert feedback, their performance improves. Cognitive modeling -based decision support isn't meant to supplant non-experts, but rather to give them immediate expert-based feedback, so as to help them make better decisions in the early trials, and reach expert-level performance at a faster rate than otherwise would have happened.

Moreover, the projected proportion of missed threats for the human-agent team, 9.6%, is lower than it would be either for the non-expert humans, 17.5%, or for standalone SDL, 20.4%. Thus, if we cared more about missed threats than false alarms, as is often the case in cyber, human-agent teaming is the ultimate option in this paradigm.

To be clear, SDL missed-threat rate can be decreased at the cost of a higher false alarm rate via a different reward structure (it is the case that the conservative-rational SDL framework includes a reinforcement learning component that is sensitive to the reward structure). However, if it was the case that the human analyst had a high degree of trust in SDL-based alert highlights, the human-agent team elevated alerts would be a superset[11] of those that would have been elevated by the human and a superset of a large proportion of SDL-elevated alerts.

We would be remiss not to point out that a decision aid will cease to be helpful without a degree of trust from the human analyst. We project that SDL-generated cognitive models of expert analysts will impart a high degree of trust for at least two reasons – (1) model-based suggestions promise to greatly improve overall non-expert performance, and (2) according to Veksler and Buchler (2019), SDL promises to be a more transparent technique than DL, one that is able to provide some explainability for each of its suggestions. The full extent to which such performance improvement and transparency may aid in establishing trust with human participants remains a topic for future research.

Overall, we find these results promising, and argue strongly that cognitive modeling can be highly useful for learning from expert analyst preferences and simulating expert-like decisions in the context of cyber support or training.

3. DYNAMIC COGNITIVE MODELS OF ATTACKER PREFERENCES

Cybersecurity is, at its core, a fundamentally adversarial paradigm. It comprises a repeated cycle where cyber defense specialists attempt to predict potential attack paths, and a cyber attacker attempts to pick an attack path to overcome the potential defense strategies. This formalism lends itself well to Game Theory (GT) -based approaches for repeated security games.

Indeed, GT-based software has been successfully applied in real-world security contexts, including airport security, coast guard, police, and anti-poaching (animal preservation) efforts (Tambe et al., 2014), providing much-needed evidence that theory based on small toy problems scales to real-world asymmetric[12] security contexts. Real world security decision aids go beyond normative game theory (picking some optimal mix of actions assuming a perfectly rational opponent), and attempt to include attacker *subjective* utilities in the equation. Recent research has shown that GT approaches to defense can be improved by relaxing the assumption of human optimal behavior and updating assumed attacker subjective utilities based on known attacker actions and feedback (Abbasi et al., 2015; Kar et al., 2015; Cooney et al., 2019; Cranford et al., 2019).

Veksler and Buchler (2016) provide simulation predictions showing that cognitive modeling -based approaches can thwart 10–50% more attacks than normative GT approaches. Specifically, they describe how Reinforcement Learning (RL)-type models may be tuned to individual attacker's subjective preferences and learning abilities via *model tracing* and *dynamic parameter fitting*. The model tracing technique makes use of boot-strapping to force-feed the participant's current experiences to the cognitive model. That is, if the participant and the model were to choose different strategies, model actions would be overwritten with participant actions in the model's memory. This method was employed in computerized instructional aids, "cognitive tutors," for students learning high school math (Anderson et al., 1995). Dynamic parameter fitting is used to adjust model parameters based on known data points, so as to make better individual predictions for future behavior. That is, if there is a free parameter in the model (such as the learning-rate parameter in the Veksler and Buchler, 2016, simulations), a range of values for this parameter are plugged into the model, and the value that best fits individual's recorded behavior is then retained for predicting their future behavior. This method was employed to predict performance of individual F-16 pilot teams (Jastrzembski et al., 2009) and is employed in software that predicts optimal training schedules based on individual performance histories (Jastrzembski et al., 2014).

The RL model used in the Veksler and Buchler (2016) simulations, as well as in the experiments described below, is based on the ACT-R utility learning mechanism (Fu and Anderson, 2006; Anderson, 2007). The model-tracing RL assumes a human attacker's action preferences will change based on their experience. For example, if the attacker chooses $A1$ and happens to lose, they will be less likely to choose $A1$ in future attacks, regardless of whether $A1$ is ultimately a good choice. Conversely, if the attacker chooses $A1$ and happens to win, they will be more likely to choose $A1$ in future attacks, regardless of whether $A1$ is ultimately a poor choice. More formally, after performing some action, A, the expected utility of this action, U_A, is incremented by the following term:

$$\Delta U_A = \alpha(R - U_A), \tag{1}$$

where α is the learning rate, and R is the value of the feedback (e.g., success/failure, reward/punishment).

Experiments 2a and 2b below attempt to validate Veksler and Buchler (2016) predictions and provide an in-depth analysis as

[11] Set A is a superset of B, or equivalently set B is a subset of A, if B is contained in A.

[12] Attacker-defender dynamics are naturally asymmetric, where the attacker is much less limited in their methods than the defender in their countermeasures.

to overall and individual effectiveness of using predictive models to pick strategies against human attackers. The simulations employed an abstract security game paradigm where the defender and attacker each had four potential strategies to choose from (payoff table showing attack success probabilities displayed as **Table 3**). Although the security game setup is abstract enough that it can fit any security context[13], we would argue that it is particularly relevant in the context of cyber security, as this is a domain where the state of the task-environment and human actions are immediately recordable.

3.1. Experiment 2a

Experiment 2a was designed to validate Veksler and Buchler (2016) simulation predictions for how the use of model tracing (MT) and model tracing in combination with dynamic parameter fitting (MT++) can improve upon game theory-based Fixed-strategy[14] and optimal Mixed-strategy[15] approaches. That is, this experiment comprised a repeated game scenario where each human participant played the role of an attacker, playing against some computational agent defender. The four between-subject conditions in this experiment corresponded to four types of computational agent defenders that human attackers were playing against: Fixed, Mixed, MT, and MT++.

This study was performed online, using Amazon Mechanical Turk to recruit adult residents of the United States for pay. We recruited 40 participants per condition.

Participants were payed 50 cents, plus five cents per win (maximum total of $3.00), and were notified as to this pay structure prior to the study. Experiment instructions were randomly altered to employ one of two contexts corresponding to cybersecurity and physical security games (see **Table 2**).

Each human participant played 50 games (i.e., trials) against their respective opponent. On each trial participant made two binary choices (see experiment instructions in **Table 2**), thus choosing among four potential attack types for that game. The computational agent also chose among four potential actions. Just as in the predictive simulations described by Veksler and Buchler (2016), the probability of a successful attack was based on the strategy selections of both players, as shown in **Table 3**. After a participant made their choices, they were alerted as to whether the attack was successful or not, and then the next game instance began.

3.1.1. Results

The top of **Figure 4** shows the performance of the different computational agents against the human attackers from Experiment 2a over the 50 trials. Performance is measured by whether or not the computational agent selected the response that maximized its probability of winning against the attacker, referred to as the optimal response[16]. As there are four possible responses, and a single response that maximizes the probability of winning, random play would result in selecting the optimal response 25% of the time.

To test for differences between the computational agent strategies and random play, we ran a mixed effect logistic regression using whether the computational agent selected an optimal response as the dependent variable, the type of computational agent as fixed effects, and the participant against which it played as random effects, with a fixed intercept of log(1/3), i.e., 25%. The logistic regression addresses the binary nature of our outcome measure; the random effects account for the multiple measures coming from the same participant; and the fixed intercept provides comparisons with our baseline of interest (random play). Although both computer and human agents may learn over time, the regression focuses on "aggregate" performance over the 50 trials and does not include a covariate for trial. The fixed strategy does significantly worse than random play, whereas both model tracer strategies do significantly better than random play. As the mixed strategy is effectively random, it did not significantly differ from the random play baseline, as expected.

While both model tracer strategies did better than random play, the effect sizes (approx. +7 percentage points for MT, and +5 percentage points for MT++) were relatively modest. We speculate (and test this speculation in Experiment 2b) that participants may have been able to adapt to the model tracer strategies, learning to be more "unpredictable." In *post-hoc* analysis, we evaluated how well MT and MT++ models could predict human decisions across all four treatments. Both MT and MT++ models appear better at predicting participant behavior in the fixed and mixed conditions than in the model tracing conditions (see **Figure 5**). There also appears to be a potential interaction in that each model tracer may be better at predicting behavior in the condition for the *other* model tracer (MT++ predicts a higher rate of decisions where human participants are playing against MT than against MT++, and MT predicts a higher rate of decisions where human participants are playing against MT++ than against MT). In combination, these findings are consistent with participants adapting to the model tracer agents, making themselves less predictable.

In analyzing the last 20 trials of the MT++ condition we find that 40–60% of the decisions were still predictable for ten out of the 40 participants (25%); 25 of the remaining participants were predictable 15–35% of the time (25% is chance), and the final five participants in this condition were predictable for $\leq$ 10% of their last 20 decisions (see **Figure 6**). The chances of the predicted choice being avoided 20 times in a row, as one of the participants managed to do, are about 3:1,000. The indication is that these individuals can predict what the predictive agent will predict, and do the opposite. Thus, it seems that keeping predictive

[13] In the experiments below we employed both cyber and physical security contexts, and found no significant difference in performance between the two contexts, $p > 0.2$.

[14] In game theory a fixed strategy— always employing the same action path—is often employed to establish a baseline level of performance.

[15] In game theory an optimal mixed strategy is one where strategies are selected randomly from a specific distribution that is tuned so as to leave the opponent no preferred choices.

[16] Maximizing the probability of winning is actually what the defender agents are designed to do; so optimal response is, in effect, a direct measure of success for the defender agent. Of course, when different attacker action paths are aimed at different targets of unequal value, optimal response is one that maximizes the likelihood of preventing an attack weighed by target value.

TABLE 2 | Experiment 2 instructions.

Cyber game	IED game
In this study you will play multiple rounds of the Cyber game. The game has two sides the Blue Force and the Red Force. The Blue Force aims to protect sensitive data. The Red Force aims to hack into the Blue Force computer network and steal the protected data. The Blue Force player will be controlled by the computer. In each round Blue Force will pick its strategy regarding which parts of the network to scan for intrusions, and how to perform those scans. You are assigned the role of Commander of the Red Force. In each round you will make 2 choices to select Red Force strategy: • Whether to focus your team's attack on the main server, or to distribute the attack over multiple servers. • Whether to scan for vulnerabilities intermittently (safer, less likely that the scan will be detected), or to scan continuously (faster). At the end of each round you will be notified whether you win the round (i.e., data acquired), or lose the round (i.e., no data was acquired).	In this study you will play multiple rounds of the IED game. The game has two sides the Blue Force and the Red Force. The Blue Force aims to deliver aid to a village. The Red Force aims to block the Blue Force form getting to the village by planting Improvised Explosive Devices (IEDs) along village routes. The Blue Force player will be controlled by the computer. In each round Blue Force will pick its strategy regarding which roads to use to get to the village, and whether to deliver aid fast, or to use more caution. You are assigned the role of Commander of the Red Force. In each round you will make 2 choices to select Red Force strategy: • Whether to plant all IEDs along the main road, or split your IEDs up and plant them along multiple roads leading into your village. • Whether to use stealth movements (safer), or to go without stealth (faster). At the end of each round you will be notified whether you win the round (i.e., Blue Force is defeated), or lose the round (i.e., Blue Force succeeds in their mission).

TABLE 3 | Payoffs for the attacker (probability of successful attack) in a security game used for Veksler and Buchler (2016) simulation predictions, as well as in Experiments 2a and 2b.

		Attacker			
		A1	**A2**	**A3**	**A4**
Defender	D1	0.15	0.45	0.50	0.90
	D2	0.55	0.10	0.90	0.45
	D3	0.50	0.90	0.15	0.45
	D4	0.90	0.50	0.50	0.10

There are four possible actions for the defender {D1,D2,D3,D4}, and four possible actions for the attacker {A1,A2,A3,A4}.

abilities hidden until some critical juncture would increase model ability to thwart attacks at said juncture. Experiment 2b examines this hypothesis.

3.2. Experiment 2b

Experiment 2a results suggest that human opponent decisions are easier to predict when said opponent does not know that they are playing against a predictive agent. In other words, if we can predict attacker actions, but then withhold this predictive ability, we can achieve high levels of success at some later critical point in time. Experiment 2b was designed to validate this hypothesis.

The methods in Experiment 2b are the same as those in Experiment 2a, with the exception of defender agent types. Specifically, this study includes three conditions, corresponding to three new agent types. The first of the agents is the MT++ agent from Experiment 2a, but it plays a fixed strategy for the first 30 of the 50 games (Fixed-MT++). The second of the agents is the MT++ agent from Experiment 2a, but it plays a mixed strategy for the first 30 of the 50 games (Mixed-MT++). The third agent, added as a control condition, plays a mixed strategy for the first 30 games, and plays a fixed strategy for the remaining 20 (Mixed-Fixed).

3.2.1. Results

The bottom of **Figure 4** shows the performance of the new multiple strategy computational agents against the human attackers from Experiment 2b over the 50 trials. As in Experiment 2a, performance is measured by optimal response.

To test for differences between the computational agents and random play and to look at performance across the distinct strategy phases, we ran a mixed effect logistic regression using optimal response as the dependent variable, the type of computational agent interacted with whether the trials were in the first segment (first 30 trials) or in the second segment (last 20 trials) of the game as fixed effects, and the participant as random effects, with a fixed intercept of log(1/3). We reproduce our results from Experiment 2a in both portions of the game for each strategy—finding that the fixed strategy is significantly less likely to select the optimal response, and both model-tracing strategies are significantly more likely to select the optimal response relative to random play in both portions of the game. The mixed strategy is not significantly different from random play. In addition, in our multiple strategy agents, we find that the Fixed-MT++ agent is significantly less likely to select the optimal response in the first 30, and more likely to select the optimal response in the last 20 compared to random play; the Mixed-MT++ is not significantly different from random play in the first 30, but significantly more likely to select the optimal response in the last 20; and the Mixed-Fixed is not significantly different from random play in the first 30, but significantly less likely to select the optimal response in the last 20.

To look more closely at how performance changes in the multiple strategy agents, and to compare the multiple strategy MT++ agents to baselines, we (1) test contrasts of agent performance between the first and second segments of the multiple strategy agents, and (2) test contrasts of MT++ performance in the last 20 trials of the multiple strategy MT++ agents and MT++ performance in the first 30 and last 20 trials of the MT++ strategy from Experiment 2a. These contrasts are tested as a single family using the mvt adjustment from the R lsmeans package.

In terms of performance changes, results confirm that between the first 30 and the last 20 trials, performance significantly increases in the Fixed-MT++ agent (**Figure 4** bottom-row, left), $p < 0.001$; significantly increases in the Mixed-MT++ agent

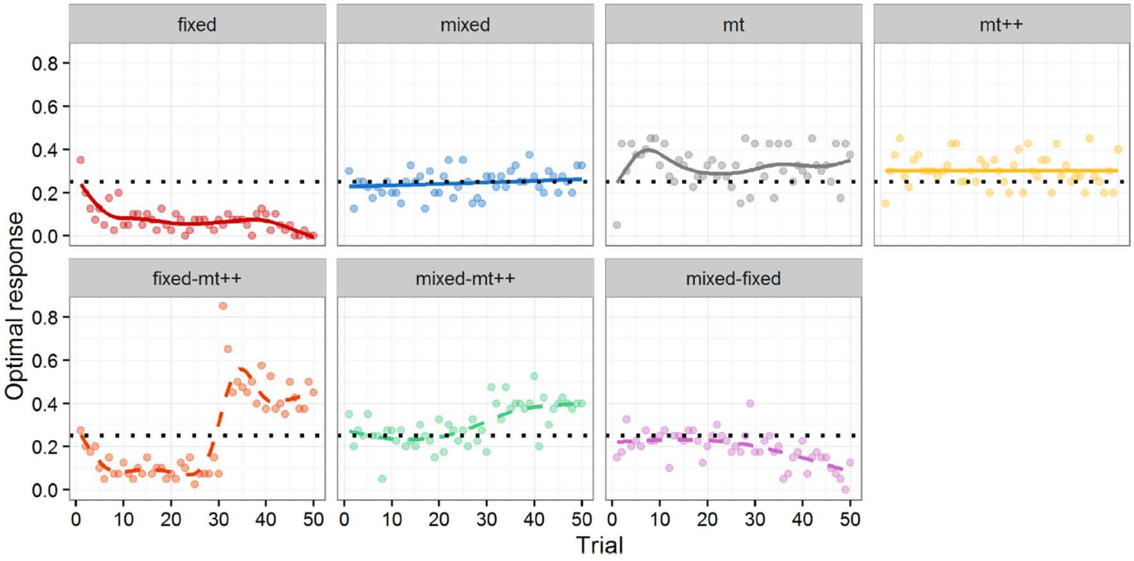


FIGURE 4 | Optimal response by trial for computational agent. Markers represent percentage of optimal response per trial. Lines represent a LOESS curve fit on data from individual participants. Solid lines **(top)** are from Experiment 2a and dashed lines **(bottom)** from Experiment 2b. Experiment 2b results all include a shift (i.e., a switch) from one type of opponent to another after 30 trials (e.g., in the *fixed-mt++* condition, human participants play against a Fixed-strategy agent for 30 games, and then against MT++ agent for the last 20 games). Dotted horizontal line indicates expected performance from random play.

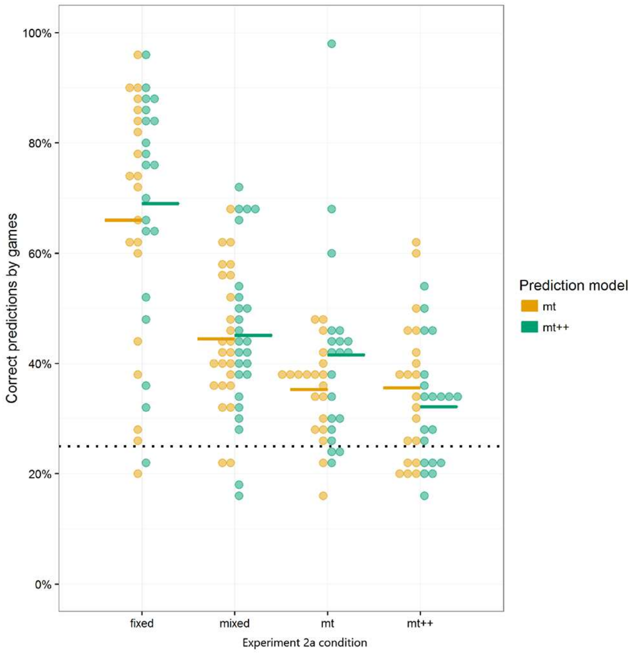


FIGURE 5 | MT and MT++ *post-hoc* predictions of participant behavior across all 4 treatments. Dots represent individual participants and dashes represent averages. Horizontal dotted lines represent expected percent of correct predictions from random guesses.

(**Figure 4** bottom-row, middle), $p < 0.001$; and significantly decreases in the Mixed-Fixed agent (**Figure 4** bottom-row, right), $p < 0.001$.

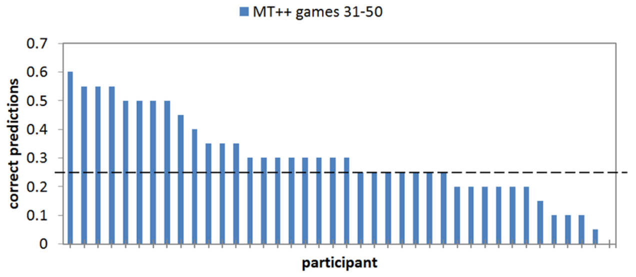


FIGURE 6 | Proportion of correct predictions of individual human participant choices in the last 20 games of the MT++ condition. Horizontal dashed line represent expected percent of correct predictions from random guesses.

In terms of relative performance, we find that when the Fixed-MT++ agent shifts to MT++ in the last 20 trials, it outperforms the normal MT++ agent in both its first 30 trials, $p = <0.001$, and its last 20 trials, $p < 0.001$; and outperforms the Mixed-MT++ agent in the last 20 trials, $p = 0.048$. When the Mixed-MT++ agent shifts to MT++ for the last 20 trials, it outperforms the normal MT++ agent in both its first 30 trials, $p = 0.006$, and the last 20 trials, $p = 0.008$.

In conjunction with the analysis of changes in performance, individual performance in the first and second segments of each game was plotted in **Figure 7**. *Post-hoc* exploratory analysis suggests additional differences between the performance of the strategies. Notably, we see a positive correlation between first and second segment behavior in the Fixed and MT agents, which suggests an element of "skill" – either on the part of the human

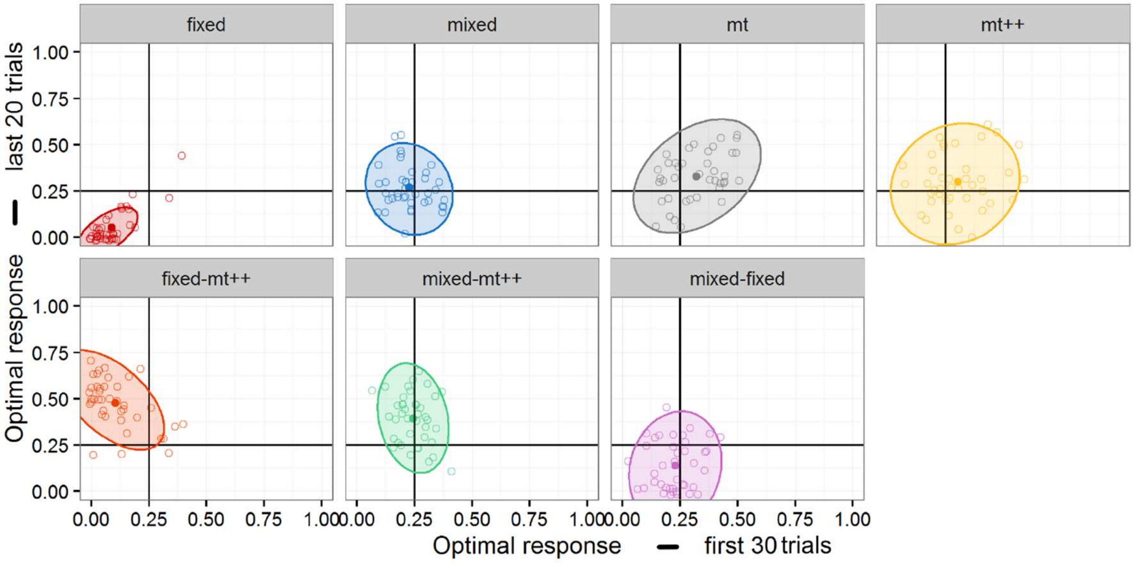


FIGURE 7 | Individual performance in first and second segments by each agent. Ellipses represent 95% confidence.

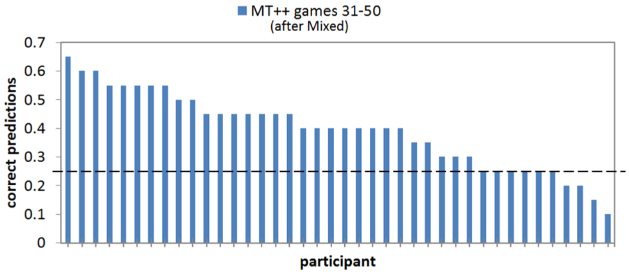


FIGURE 8 | Proportion of correct predictions of individual human participant choices in the last 20 games of the MT++ condition. Horizontal dashed line represents expected proportion of correct predictions from random guesses.

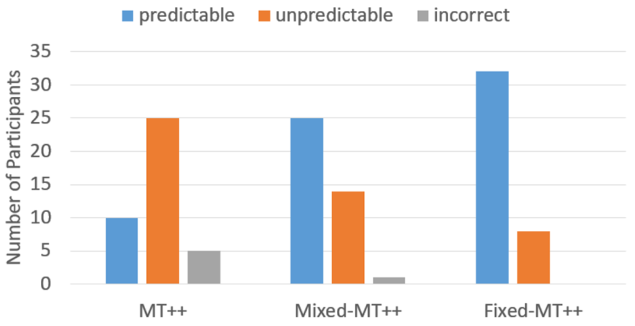


FIGURE 9 | Number of participants whose choices were predictable, unpredictable, or incorrect playing against MT++ in the last 20 trials of the experiment. Participants were labeled as *predictable* if more than 40% of their decisions were correctly predicted by the model, *unpredictable* if their decisions were predicted between 15 and 35% (25% is chance), and *incorrect* if their decisions were the opposite of what was predicted.

player or the computational agent. In contrast, neither the Mixed nor MT++ agents show any correlation, meaning that people (and computational agents) that perform relatively better in the first segment are no more likely to perform better in the second segment. This suggests that, while the MT and MT++ strategies may have similar optimal response rates as suggested by Experiment 2a, there are nonetheless differences in the way these strategies interact with their opponents.

A second relationship of interest can be noted in the second row of **Figure** 7. In particular, the Fixed-MT++ agent shows a negative correlation between first and second segments whereas the Mixed-MT++ shows a weak one—if any. One possible interpretation of this negative correlation is that, the switching strategy may be particularly effective if the opponents are able to adopt some counter-strategy in the initial trials.

When we break down the results by individual predictability we find that keeping predictive abilities hidden greatly mitigates attacker ability to adapt and become less predictable. For example, in the last 20 games of the Mixed-MT++ condition, 40–65% of the choices were predicted for 25 of the participants (62.5%), 15–35% of the choices were predicted for fourteen of the participants, and only one individual was predictable on 10% of their choices (see **Figure 8**). The results from the last 20 trials of the Fixed-MT++ condition are better still, with 32 individuals being predictable at 40–70%, and the remaining 8 at 20–35%. The overall predictability of attackers playing against MT++ the entire time (from Experiment 2a) and MT++ after switching from Mixed and Fixed strategies are shown in **Figure 9**.

3.3. Discussion

Results from these experiments confirm the general prediction that cognitive modeling techniques can be more effective than normative GT in the context of predicting attacker decisions. However, the average advantage of cognitive modeling over GT seems to be greatly diminished when attackers realize they are playing against a predictive agent. That is, when human players know that they are matched against a predictive agent, their play changes and becomes less predictable. However, this does not mean that all participants learn to play [pseudo-]randomly against predictive agents. Rather, some individual game-play remained predictable, some looked more like chance play, and some of the individuals began to predict the predictive agent, adopting a "Theory of Mind" (ToM) strategy[17]. The less "smart" of an agent human participants were matched against, the more predictable their play became, with participants that played against a Fixed-strategy agent becoming the most predictable, those playing against a Mixed-strategy agent being less so, and those playing against MT++ being the least predictable.

Ultimately, by keeping track of prediction success for each individual attacker, a defender agent should be able to ascribe the correct model to the attacker: random play, RL, or ToM. Once the correct model of the attacker is determined, the defender can choose its own appropriate strategy: Mixed strategy, a predictive strategy, or the opposite of the predictive strategy, respectively. Seemingly, once the human attacker realizes that they are playing against a predictive agent and switches to a reciprocal strategy, and the agent switches its strategy in turn, the two opponents may continue to switch their game-play continuously. However, it is not the case that this would necessarily end with Mixed strategy level of play by the two opponents, as humans are notoriously bad at being random. West and Lebiere (2001) predict that chaos-like game-play may actually be an emergent property of reciprocal and predictable human choices.

This paper only explores standalone RL as a potential cognitive model for predicting attacker choices. More sophisticated attacker models, including those based on ToM strategies, the work of West and Lebiere (2001), and models that include domain-specific knowledge, should be able to account for a greater range of attacker behavior. Attacker models can be further seeded based on types of attacker personalities, risk-tolerance, and attack-types common to specific geographic regions (e.g., Sample, 2015). Having a greater wealth of model types would be a major boon to dynamically fitting individual attacker behavior, and would result in more precise and accurate predictions of further attacker choice. In a more general sense, we argue that Cognitive Modeling as a discipline is useful for predicting individual preferences and behavior, and is thus highly relevant for real-time cybersecurity decision support.

4. SUMMARY

Prior work has argued that cognitive modeling techniques can be trained on expert data so as to provide such expertise as an aid for non-experts, and that CM-based Symbolic Deep Learning would be more useful in this endeavor than ML-based Deep Learning frameworks, especially in fields like cybersecurity where expert data is not highly abundant (Veksler and Buchler, 2019). In a similar vein, other work has made strong predictions that cognitive modeling may be useful in predicting opponent decision preferences in repeated security games, and be more useful than normative GT-based security aids, especially in fields like cybersecurity where behavior/feedback of attackers can be dynamically observed/updated (Veksler and Buchler, 2016). We presented Experiments 1 and 2 above so as to examine these predictions against human data.

Experiment 1 results revealed that CM-based SDL framework is more effective than ML-based DL framework in learning from experts and has much more potential for improving non-expert performance. The separation between SDL and DL effectiveness greatly increases as the available training data gets more sparse. Regardless of model type or training data, it is the case that a human-agent team where the human non-expert always accepts model suggestions for elevating alerts will have a higher alert hitrate than either a lone human or a lone model. Future work in this domain will focus on the topic of trust, and examining the degree to which SDL-based decision-aids will be trusted by human non-experts. We project that the overall level of performance improvement and the potential for decision-explainability that SDL-generated cognitive models can provide will create enough trust to develop highly effective teams of human-agent analysts.

Experiment 2 results revealed that model-tracing and dynamic parameter-fitting techniques can be used to continuously update cognitive models of attackers and to accurately predict a high percentage of their decisions. Results further indicate that when model predictive capabilities are hidden from the opponent, the opponent's decisions become more predictable, especially when said opponent believes they are playing against an unsophisticated defender. Our conclusion is that in adversarial repeated cybersecurity contexts cognitive models should be tuned to individual attacker's preferences, but model predictive abilities should be held hidden until some critical juncture so as to maximize effect. Future work will focus on development of more sophisticated models, such that when attackers recognize a model's predictive abilities and attempt to pivot to unforeseen strategies, the model can make a timely pivot, as well.

The experiment and simulation results presented here look promising. However, this paper presents theoretical models examined in absence of real-world confounds. Future work will focus on stress-testing these models in the context of real cybersecurity data. Although it is the case that "[t]here is nothing so useful as a good theory," (Lewin, 1951, as cited by Gray and Altmann, 2001) it is also the case that "[n]othing drives basic science better than a good applied problem" (Newell and Card, 1985, as cited by Gray and Altmann, 2001). We believe that the methods presented in this paper can be of great use for cybersecurity, but also that the applied problem of cybersecurity itself and the datasets derived in this domain can serve to refine these methods and to push them from research stages and toward production.

On a more general note, we would argue that Cognitive Science, and specifically Cognitive Modeling as a discipline, is highly relevant and holds great promise in cybersecurity and analogous domains. Models of human cognition can be automatically tuned to either defender or attacker preferences, and such models can then be used in simulations, training, and

[17] The concept of Theory of Mind refers to one's ability to infer others' beliefs and intentions (e.g., Hiatt and Trafton, 2010).

decision aids. Whereas network/software vulnerabilities change constantly, the fundamentals of human learning and decision-making principles remain the same. In taking advantage of established and emerging cognitive and behavioral research and technology we can vastly improve our overall long-term network safety.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by U.S. Army Research Laboratory. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

FUNDING

Project funding from the Office of the Under-Secretary of Defense for Research & Engineering (OUSD(R&E)) Cyber Technologies Program. CG, MY, and CLe were funded by the Army Research Laboratory under Cooperative Agreement W911NF-13-2-0045 (ARL Cyber Security CRA). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The authors thank research assistants in the Dynamic Decision Making Laboratory, Carnegie Mellon University, for their help with data collection.

REFERENCES

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., et al. (2016). TensorFlow: large-scale machine learning on heterogeneous distributed systems. *arXiv [Preprint].* arXiv:1603.04467.

Abbasi, Y. D., Short, M., Sinha, A., Sintov, N., Zhang, C., and Tambe, M. (2015). "Human adversaries in opportunistic crime security games: evaluating competing bounded rationality models," in *Proceedings of the Third Annual Conference on Advances in Cognitive Systems ACS.* (Atlanta, GA) 2.

Anderson, J. R. (1993). *Rules of the Mind*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Anderson, J. R. (2007). *How Can the Human Mind Occur in the Physical Universe*? Oxford; New York, NY: Oxford University Press.

Anderson, J. R., Corbett, A. T., Koedinger, K. R., and Pelletier, R. (1995). Cognitive tutors: lessons learned. *J. Learn. Sci.* 4:167207. doi: 10.1207/s15327809jls0402_2

Anderson, J. R., and Lebiere, C. (1998). *The Atomic Components of Thought*. Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Cooney, S., Vayanos, P., Nguyen, T. H., Gonzalez, C., Lebiere, C., Cranford, E. A., et al. (2019). "Warning time: optimizing strategic signaling for security against boundedly rational adversaries," in *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems* (Montreal, QC: International Foundation for Autonomous Agents and Multiagent Systems), 1892–1894.

Cranford, E. A., Gonzalez, C., Aggarwal, P., Cooney, S., Tambe, M., and Lebiere, C. (2019). "Towards personalized deceptive signaling for cyber defense using cognitive models," in *17th Annual Meeting of the International Conference on Cognitive Modelling (ICCM 2019)* (Madison, WI).

D'Amico, A., and Whitley, K. (2008). "The real work of computer network defense analysts," in *VizSEC 2007*, eds J. R. Goodall, G. Conti, and K. L. Ma (Berlin; Heidelberg: Springer), 19–37. doi: 10.1007/978-3-540-78 243-8_2

Dutra, A. R. A., Garcez, A., and D'Avila Garcez, A. S. (2017). "A comparison between deepQ-networks and deep symbolic reinforcement learning," in *Proceedings of the Twelfth International Workshop on Neural-Symbolic Learning and Reasoning 2017* (London).

Feigenbaum, E., and Simon, H. (1984). EPAM-like models of recognition and learning. *Cogn. Sci.* 8, 305–336. doi: 10.1207/s15516709cog0804_1

Friesen, M. A., White, S. V., and Byers, J. F. (2008). "Chapter 34: Handoffs: implications for nurses," in *Patient Safety and Quality: An Evidence-Based Handbook for Nurses*, ed R. G. Hughes (Rockville, MD: Agency for Healthcare Research and Quality), 285–332.

Fu, W. T., and Anderson, J. R. (2006). From recurrent choice to skilled learning: a reinforcement learning model. *J. Exp. Psychol. Gen.* 135, 184–206. doi: 10.1037/0096-3445.135.2.184

Garcez, A., Dutra, A. R. R., and Alonso, E. (2018). Towards symbolic reinforcement learning with common sense. *arXiv [Preprint]*. Retrieved from: http://arxiv.org/abs/1804.08597

Gluck, M. A., and Bower, G. H. (1988). From conditioning to category learning - an adaptive network model. *J. Exp. Psychol. Gen.* 117, 227–247. doi: 10.1037/0096-3445.117.3.227

Gobet, F. (1998). Expert memory: a comparison of four theories. *Cognition* 66, 115–152. doi: 10.1016/S0010-0277(98)00020-1

Gonzalez, C., Ben-Asher, N., Oltramari, A., and Lebiere, C. (2014). "Cognition and Technology," in *Cyber Defense and Situational Awareness*. Advances in Information Security, Vol 62, eds A. Kott, C. Wang, and R. Erbacher (Cham: Springer), 93–117. doi: 10.1007/978-3-319-11391-3_6

Gray, W. D., and Altmann, E. M. (2001). "Cognitive modeling and human-computer interaction," in *International Encyclopedia of Ergonomics and Human Factors*, Vol. 1, ed W. Karwowski (New York, NY: Taylor & Francis, Ltd.), 387–391.

Hiatt, L. M., and Trafton, G. J. (2010). "A cognitive model of theory of mind," in *Proceedings of the 10th International Conference on Cognitive Modeling, ICCM 2010* (Philadelphia, PA).

Jastrzembski, T. S., Gluck, K. A., and Rodgers, S. (2009). "Improving military readiness: a state-of-the-art cognitive tool to predict performance and optimize training effectiveness," in *The Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC)* (Orlando, FL).

Jastrzembski, T. S., Rodgers, S. M., Gluck, K. A., and Krusmark, M. A. (2014). Predictive Performance Optimizer. U.S. Patent No. 8,568,145. Washington, DC: U.S. Patent and Trademark Office.

Kar, D., Fang, F., Fave, F. D., Sintov, N., and Tambe, M. (2015). "A game of thrones: when human behavior models compete in repeated Stackelberg security games," in *2015 International Conference on Autonomous Agents and Multiagent Systems* (Istambul: International Foundation for Autonomous Agents and Multiagent Systems), 1381–1390.

Lewin, K. (1951). *Field Theory in Social Science.* New York, NY: Harper Row.

Newell, A., and Card, S. K. (1985). The prospects for psychological science in human-computer interaction. *Human Comput. Interact.* 1, 209–242.

Oltsik, J. (2019). *The Life and Times of Cybersecurity Professionals*. Technical report, Enterprise Strategy Group (ESG).

Rusk, N. (2015). Deep learning. *Nat. Methods.* 13:35. doi: 10.1038/nmeth.3707

Sample, C. (2015). *Cyber + Culture Early Warning Study*. Technical report, CERT.

Stimpfel, A. W., Sloane, D. M., and Aiken, L. H. (2012). The longer the shifts for hospital nurses, the higher the levels of burnout and patient dissatisfaction. *Health Affairs* 31, 2501–2509. doi: 10.1377/hlthaff.2011.1377

Swets, J. A. (1964). *Signal Detection and Recognition by Human Observers, 1st Edn*. New York, NY: Wiley. doi: 10.1037/e444572004-001

Swets, J. A. (1996). *Signal Detection Theory and ROC Analysis in Psychology and Diagnostics: Collected Papers*. New York, NY: Psychology Press.

Tambe, M., Jiang, A. X., An, B., and Jain, M. (2014). "Computational game theory for security: progress and challenges," in *AAAI Spring Symposium on Applied Computational Game Theory* (Palo Alto, CA). doi: 10.2197/ipsjjip.22.176

Veksler, V. D., and Buchler, N. (2016). "Know your enemy: applying cognitive modeling in security domain," in *38th Annual Conference of the Cognitive Science Society* (Philadelphia, PA).

Veksler, V. D., and Buchler, N. (2019). "Cognitive modeling with symbolic deep learning," in *17th Annual Meeting of the International Conference on Cognitive Modelling (ICCM 2019)* (Montreal, QC).

Veksler, V. D., Buchler, N., Hoffman, B. E., Cassenti, D. N., Sample, C., and Sugrim, S. (2018). Simulations in cyber-security: a review of cognitive modeling of network attackers, defenders, and users. *Front. Psychol*. 9:691. doi: 10.3389/fpsyg.2018.00691

Veksler, V. D., Gluck, K. A., Myers, C. W., Harris, J., and Mielke, T. (2014). Alleviating the curse of dimensionality–A psychologically-inspired approach. *Biol. Inspired Cogn. Architect*. 10, 51–60. doi: 10.1016/j.bica.2014.11.007

West, R. L., and Lebiere, C. (2001). Simple games as dynamic, coupled systems: randomness and other emergent properties. *Cogn. Syst. Res.* 1, 221–239. doi: 10.1016/S1389-0417(00)00014-0

Zhang, Q., and Sornette, D. (2017). Learning like humans with Deep Symbolic Networks. *arXiv [Preprint]*. Retrieved from: http://arxiv.org/abs/1707.03377

Conflict of Interest: VV was employed by the company DCS Corp.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Psychology

OPINION
published: 10 July 2020
doi: 10.3389/fpsyg.2020.01390

In Defence of the Human Factor

Ciarán Mc Mahon[1,2]*

[1] *Institute of Cyber Security, Dublin, Ireland,* [2] *School of Psychology, University College Dublin, Dublin, Ireland*

Keywords: cybersecurity, human factors, software, psychology, theory

OPEN ACCESS

Edited by:
Richard Keith Wortley,
University College London,
United Kingdom

Reviewed by:
Chris Baber,
University of Birmingham,
United Kingdom

****Correspondence:***
Ciarán Mc Mahon
ciaran@instituteofcybersecurity.com

Specialty section:
This article was submitted to
Cognition,
a section of the journal
Frontiers in Psychology

Received: *17 January 2020*
Accepted: *25 May 2020*
Published: *10 July 2020*

Citation:
Mc Mahon C (2020) In Defence of the
Human Factor.
Front. Psychol. 11:1390.
doi: 10.3389/fpsyg.2020.01390

INTRODUCTION

A trope that has long dominated cybersecurity is the idea that "humans are the weakest link." While its intellectual origins predate the industry by several decades, if not centuries, for our present purposes we need go back no further than the beginning of this millennium. It seems to have started with Schneier (2000), and continued with Mitnick and Simon (2002). Since then, cybersecurity discourse has been awash with this cliché.

In his book, Schneier (2000) discusses the idea of perfect computer security. Imagine a flawless computer, with strong cryptography and secure protocols. Even though it would be difficult, suppose it is operational. Unfortunately, it isn't secure, because sooner or later it will have to interact with a user, and "this interaction is the biggest risk of them all. People often represent the weakest link in the security chain and are chronically responsible for the failure of security systems" (Schneier, 2000, p. 149). And while Mitnick and Simon (2002) begins in a different tone, his point is essentially the same. Talking about home security, and how people install locks in order to feel safe, he says no matter what is put in place, the home remains essentially vulnerable, because "the human factor is truly security's weakest link." Schneier's and Mitnicks' influences are such that this phrase developed significant currency in information security circles, though it was likely an already common trope in physical security discourse.

"The human factor is the weakest link in cybersecurity" has acquired the status of a thought-terminating cliché, and its continued popularity is restraining the intellectual development of this field. It should be retired as an immediate concern.

But at present, cybersecurity is utterly soaked in this idea. It features prominently in security awareness blogs (Spitzner, 2012), IT industry publications (Rossi, 2015; Wright, 2016), media outlets (Vishwanath, 2016), and even Oxford University Press monographs (Singer and Friedman, 2014). Recently, at a government-sponsored event in Ireland, an afternoon panel was titled "Cybersecurity: Defending the weakest link" (Dublin Digital Summit, 2019). As such, this negative characterisation of human nature shows no sign of waning.

Notably, some scholars pushed back from the very outset (e.g., Sasse et al., 2001) but these voices have been rare. In contrast, a vast amount of literature explicitly advocated for it: in the context of airport (Schwaninger, 2006) and mobile security (Lau, 2017); systematic reviews (Mahfuth et al., 2017), cyberpsychology (Wiederhold, 2014), social networking (Lehrman, 2010)—and many more. These citations are only those which mention the phrase overtly: a more detailed reading of the literature would almost certainly expose the "human factor is the weakest link in cybersecurity" as one of the premises on which information security science's current paradigm is based (Kuhn, 1962).

Breaking the Chain

Let us scrutinise this trope dispassionately. Suppose that information security is effectively analogised as a chain of some sort, composed of links, and one of those links is the "human factor." What is the nature of this chain, and what are its other components? I won't stretch the analogy any further than is intended by its proponents. But I don't think it unreasonable to deduce that this chain is intended to be protecting the assets, information and finances of some organisation.

Apart from the "human factor," this chain comprises technical, physical, or similar synthetic links. Crucially, I presume that those who say that the "human factor is the weakest link in cybersecurity" do not have the engineers of those links in mind. No, it is clear that they are pointing toward the humans who use those links, not their creators.

What we are supposed to read from this phrase is actually "end users are the weakest link"—with the obvious corollary being that the other links—networks, software, applications—are much stronger and more secure. Computers don't make mistakes, people do.

But can this really hold up? Are the other links in the security chain really stronger? In a much-shared opinion piece for The Message, well-known internet essayist Norton (2014) argued that "Everything is broken." Putting it bluntly, she says: "It's hard to explain to regular people how much technology barely works, how much the infrastructure of our lives is held together by the IT equivalent of baling wire. Computers, and computing, are broken."

Update of the Art

The reality of the other links in the cybersecurity chain are is best illustrated by examining the current state of software updating. Take mobile operating systems. Between 1 January and 31 December 2019, Apple released ∼20 security updates to its most recent versions (i.e., 12 and 13) of its mobile operating system, iOS (Apple Inc., 2020a). In any other sphere of consumer activity, this level of patching would not be tolerated. Imagine telling car owners that they must fix their car practically every fortnight if they want to keep driving it safely. And if accidents occurred in such a scenario, would we blame the stupid drivers?

In fact, iOS is noteworthy in how persistently it encourages its users to update, with repeated notifications, pop-ups and warnings. The net result that a sizeable proportion of users have installed the latest version. As of October 2019, 50% of all iOS devices are using the most recent version of the software (Apple Inc., 2020b).

On the other hand, its main competitor, the Google-owned Android, is not known for this kind of encouragement. Its most recent version, Android 10, was released in September 2019 but Google has yet to update its distribution statistics since May 2019. At that point, only 10.4% of all Android devices were running the preceding most up-to-date version, known as Pie (Android Developers, 2020). Hence, presumably a much smaller percentage are using the newer Android 10. This sorry state of affairs was such that it was for a time investigated by the both the Federal Trade Commission and Federal Communication Commission of the United States (Rossi, 2015).

These are far from the worst examples—the soon-to-be deprecated Adobe Flash Player pushed out an extraordinary number of updates over the course of its history—on occasions pushing out three updates within a month (Adobe, 2020). How are users supposed to keep up? Another example some may recall is the problematic release of the Windows 8 operating system. While usually the release of such a massive piece of work follows several years of careful engineering, Windows 8 was quickly beset by a host of user-reported difficulties. Hence, it was succeeded is less than a year by Windows 8.1—as a free update (LeBlanc, 2013).

This is the real problem in information security—it's not the end users who are to blame, it's the fact that so much rickety code is being pushed out without being properly secured. But then why do we say that the "human factor is the weakest link," when the other links need constant repair?

What Is Human Error?

The answer is simply that blaming the end user for a breach falls into the category of "acceptable accident causes." Hollnagel and Amalberti (2001), in studying a context not dissimilar to cyber attacks, namely industrial safety, note that accidents are always found to have been clearly associated with a particular aspect or function of a system. Such an aspect or function can be corrected within accepted limits of cost and time and conforms to current "norms" for explanations.

Clearly, when we talk about breaches, the human factor fits into this framework of an acceptable cause. An individual made a mistake and they will be fired: this is what we expect to happen. Blaming an end user is an easy way of explaining what happened, rather than solving the much more difficult and costly problem of the patchy state of networked computing.

We need more of a systems approach to the human factor in cybersecurity *á la* Reason (2000). In a classic paper on mishaps in medical practice, Reason outlined a "Swiss cheese" model of error, where safeguards from harm are imagined as individual slices of cheese, each with its own holes or weaknesses. Occasionally, these line up, allowing an "accident trajectory" to form. Evidently, when "everything is broken" in information technology, such trajectories can occur frequently.

Hence, Hollnagel (1983) argues that human error is a meaningless concept. It makes no sense to castigate individuals for doing something which yesterday was correct, but today is wrong. Take phishing, for example. Every day the average office worker clicks on probably hundreds of hyperlinks as part of their job, whether searching the internet or opening emails. Then 1 day, they click on the wrong one, and suddenly they're the cause of a malware infection.

But not only is the end user the end point in a breach trajectory over which they have little control, they are also at the mercy of heavily automated systems. Because software detection of phishing attacks is improving, end users are less exposed to them. Hence, they learn less about how to recognise such risky emails and are less prepared for dealing with them when they do arrive. Calling to mind Bainbridge (1983) "irony of automation," the stupid human has largely been designed out of how the system handles risk. Consequently, it is surely unfair to blame them when they become the end point of a breach trajectory.

Stop Blaming the Victim

However, that's not the only reason we shouldn't say "the human factor is the weakest link in cybersecurity"—there are important psychological factors too. Firstly, blaming the user for compromises can be seen as a form of victim blaming. Cross (2015) argues that discourse on online fraud is based on idea of greedy or gullible victims and does not take into account level of deception and sophisticated targeting that is behind it. More crucially, this victim-blaming discourse isolates victims and impacts their ability to warn others.

Secondly, in an organisational context the idea that the human factor is a "weak link," is often supplemented with the suggestion that it is often a harmful one too—i.e., not only causing breaches accidentally, but deliberately. However, in a study examining abusive insiders, Posey et al. (2011) show that employees who do not feel that their organisations trust them will engage in more computer abuse when new security measures are introduced.

Additionally, in a highly-cited study of organisational justice, Bulgurcu et al. (2009) demonstrate that creating a sense of procedural fairness with regard to rules and regulations is the key to effective information security management. In sum, it is important that, far from presuming that they are the "weakest link," our end users be dealt with fairly and with trust.

Finally, in a survey of 118 senior European information security professionals, only 29% of respondents could agree (or strongly agree) that "end user errors or violations are disciplined fairly and transparently, regardless of seniority" (Barker et al., 2020). If these data are reflective of organisations at large, it would seem that most of them are not governed with any real sense of justice when it comes to cybersecurity. We cannot expect end users to follow information security policy in such an environment.

CONCLUSION

I regret I have not had the chance to offer any tangible solutions in this brief overview. So, in order to help to retire this trope, here are some questions I suggest readers ask when they encounter the "human being is the weakest link" trope.

- How would we expect our colleagues to react if we were to describe them personally like this?
- What are the other links in this chain and how secure are they really?
- What breach trajectory must be created before a human being can become a weak link?
- Has the human been automated out of the system in question?
- Am I blaming the victim of a crime? Am I treating end users fairly and transparently?
- Fundamentally, why are we pushing such a negative vision of human capability? Who exactly are we serving with such a message?

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

REFERENCES

Adobe (2020). *Release Notes | Flash Player® 32 AIR® 32*. Retrieved from: https://helpx.adobe.com/flash-player/release-note/fp_32_air_32_release_notes.html#id_62973 (accessed June 2020).

Android Developers (2020). *Distribution Dashboard*. Retrieved from: https://developer.android.com/about/dashboards (accessed June 18, 2020).

Apple Inc. (2020a). *Apple Security Updates*. Retrieved from: https://support.apple.com/en-us/HT201222 (accessed June 18, 2020).

Apple Inc. (2020b). *App Store*. Retrieved from: https://developer.apple.com/support/app-store/ (accessed June 18, 2020).

Bainbridge, L. (1983). Ironies of automation. *Automatica* 19, 775–779. doi: 10.1016/0005-1098(83)90046-8

Barker, J., Davis, A., Hallas, B., and Mc Mahon, C. (2020). *Cyber Security ABCs: Delivering Awareness, Behaviours and Culture Change*. London: British Computer Society.

Bulgurcu, B., Cavusoglu, H., and Benbasat, I. (2009). "Roles of information security awareness and perceived fairness in information security policy compliance," in *15th Americas Conference on Information Systems, AMCIS 2009, Vol. 5*, eds K. E. Kendall and U. Varshney (San Francisco: AIS), 3269–3277.

Cross, C. (2015). No laughing matter: blaming the victim of online fraud. *Int. Rev. Vict.* 21, 187–204. doi: 10.1177/0269758015571471

Dublin Digital Summit (2019). *Programme*. Dublin. Retrieved from: http://digitalsummitdublin.ie/programme/ (accessed June 18, 2020).

Hollnagel, E. (1983). "Why human error is a meaningless concept," in *NATO Conference on Human Error* (Bellagio). Retrieved from: http://158.132.155.107/posh97/private/humanfactors/hollnagel.pdf (accessed June 18, 2020).

Hollnagel, E., and Amalberti, R. (2001). "The emperor's new clothes: or whatever happened to 'human error'?" in *4th International Workshop on Human Error, Safety and Systems Development* (Linköping, Linköping University).

Kuhn, T. S. (1962). *The Structure of Scientific Revolutions*. Chicago, IL: University of Chicago Press.

Lau, L. (2017). "Mobile Security," in *Mobile Security and Privacy*, eds M. H. Au and K.-K. R. Choo (Cambridge, MA: Elsevier), 57–66. doi: 10.1016/B978-0-12-804629-6.00003-1

LeBlanc, B. (2013, October 17). *Windows 8.1 Now Available!* Windows Blogs. Retrieved from: https://blogs.windows.com/windowsexperience/2013/10/17/windows-8-1-now-available/ (accessed June 18, 2020).

Lehrman, Y. (2010). The weakest link: the risks associated with social networking websites. *J. Strategic Security* 3, 63–72. doi: 10.5038/1944-0472.3.2.7

Mahfuth, A., Yussof, S., Baker, A. A., and Ali, N. (2017). "A systematic literature review: information security culture," in *2017 International Conference on Research and Innovation in Information Systems (ICRIIS)* (Langkawi, Malaysia: IEEE), 1–6. doi: 10.1109/ICRIIS.2017.8002442

Mitnick, K. D., and Simon, W. L. (2002). *The Art of Deception: Controlling the Human Element of Security*. Indianapolis, IN: John Wiley & Sons.

Norton, Q. (2014, May 20). 'Everything is broken'. *The Message*. Retrieved from: https://medium.com/message/everything-is-broken-81e5f33a24e1#.sc7pf19g3 (accessed June 18, 2020).

Posey, C., Bennett, R. J., and Roberts, T. L. (2011). Understanding the mindset of the abusive insider: an examination of insiders' causal reasoning following internal security changes. *Comput Secur.* 30, 486–497. doi: 10.1016/j.cose.2011.05.002

Reason, J. (2000). Human error: Models and management. *Br. Med. J.* 320, 768–770. doi: 10.1136/bmj.320.7237.768

Rossi, B. (2015, June 30). The human factor: top tips to strengthen the weakest link in the information security chain. *Information Age*. Retrieved from: http://www.information-age.com/technology/security/123459735/human-factor-top-tips-strengthen-weakest-link-information-security-chain (accessed June 18, 2020).

Sasse, M. A., Brostoff, S., and Weirich, D. (2001). Transforming the "weakest link" - A human/computer interaction approach to usable and effective security. *BT Technol. J.* 19, 122–131. doi: 10.1023/A:1011902718709

Schneier, B. (2000). *Secrets and Lies: Digital Security in a Networked World*. New York, NY: John Wiley & Sons.

Schwaninger, A. (2006). "Airport security human factors: from the weakest to the strongest link in airport security screening," in *Proceedings of the 4th International Aviation Security Technology Symposium* (Washington, DC), 265–270. doi: 10.13140/RG.2.1.1561.4965

Singer, P. W., and Friedman, A. (2014). *Cybersecurity: What Everyone Needs to Know*. Oxford: OUP.

Spitzner, L. (2012, September 17). This is why the human is the weakest link. *SANS Security Awareness Blog*. Retrieved from: https://www.sans.org/security-awareness-training/blog/why-human-weakest-link (accessed June 18, 2020).

Vishwanath, A. (2016, May 5). Cybersecurity's weakest link: humans. *The Conversation*. Retrieved from: https://theconversation.com/cybersecuritys-weakest-link-humans-57455 (accessed June 18, 2020).

Wiederhold, B. K. (2014). The role of psychology in enhancing cybersecurity. *Cyberpsychol. Behav. Soc. Netw.* 17, 131–132. doi: 10.1089/cyber.2014.1502

Wright, A. (2016, April 13). Humans in cyber security – the weakest link. *IT Governance*. Retrieved from: https://www.itgovernance.co.uk/blog/humans-in-cyber-security-the-weakest-link/ (accessed June 18, 2020).

Conflict of Interest: CM was employed by the company Institute of Cyber Security.

frontiers in Psychology

ORIGINAL RESEARCH
published: 28 July 2020
doi: 10.3389/fpsyg.2020.01756

Understanding Phishing Email Processing and Perceived Trustworthiness Through Eye Tracking

John McAlaney and Peter J. Hills*

Faculty of Science & Technology, Department of Psychology, Bournemouth University, Poole, United Kingdom

Social engineering attacks in the form of phishing emails represent one of the biggest risks to cybersecurity. There is a lack of research on how the common elements of phishing emails, such as the presence of misspellings and the use of urgency and threatening language, influences how the email is processed and judged by individuals. Eye tracking technology may provide insight into this. In this exploratory study a sample of 22 participants viewed a series of emails with or without indicators associated with phishing emails, whilst their eye movements were recorded using a SMI RED 500 eye-tracker. Participants were also asked to give a numerical rating of how trustworthy they deemed each email to be. Overall, it was found that participants looked more frequently at the indicators associated with phishing than would be expected by chance but spent less overall time viewing these elements than would be expected by chance. The emails that included indicators associated with phishing were rated as less trustworthy on average, with the presence of misspellings or threatening language being associated with the lowest trustworthiness ratings. In addition, it was noted that phishing indicators relating to threatening language or urgency were viewed before misspellings. However, there was no significant interaction between the trustworthiness ratings of the emails and the amount of scanning time for phishing indicators within the emails. These results suggest that there is a complex relationship between the presence of indicators associated with phishing within an email and how trustworthy that email is judged to be. This study also demonstrates that eye tracking technology is a feasible method with which to identify and record how phishing emails are processed visually by individuals, which may contribute toward the design of future mitigation approaches.

Keywords: phishing, eye tracking, social engineering, cybersecurity, email

OPEN ACCESS

Edited by:
Paul Watters,
La Trobe University, Australia

Reviewed by:
Rakesh Verma,
University of Houston, United States
Karen Renaud,
Abertay University, United Kingdom

****Correspondence:***
John McAlaney
jmcalaney@bournemouth.ac.uk

Specialty section:
This article was submitted to Cognition, a section of the journal Frontiers in Psychology

Received: *25 February 2020*
Accepted: *25 June 2020*
Published: *28 July 2020*

Citation:
McAlaney J and Hills PJ (2020) Understanding Phishing Email Processing and Perceived Trustworthiness Through Eye Tracking. Front. Psychol. 11:1756. doi: 10.3389/fpsyg.2020.01756

INTRODUCTION

Internet browsers, email systems and other socio-technical systems require input from individual users. Such systems may be designed in a way that aims to protect users and organizations from external attackers as much as is possible (Das et al., 2020). How successful they are in doing so is highly reliant on the user (Pfeffel et al., 2019). The user may not fully attend to the cues and prompts that the system provides them with to encourage them to use the system in a safe way (Miyamoto et al., 2015). Similarly, users may fail to detect and respond to potential threats, even if the system provides prominent indicators of these threats (Miyamoto et al., 2015). An example of this is phishing websites, links to which are sent to potential victims through phishing emails. Phishing emails are one form of social engineering, which refers to the use of manipulation and trickery to cause an individual to gain sensitive information or access to a system (Hadnagy, 2018). This type of attack has been described as the single biggest threat to cybersecurity (Salahdine and Kaabouch, 2019). Individuals who engage in cyber-crime do not need to possess programming or technological skills in order to be able to create phishing emails; software packages that can be used to create phishing emails can be downloaded online (McCalley et al., 2011).

This reliance on user engagement in many sociotechnical systems is potentially problematic. As is well established in psychological research people often do not fully processes all the information that is available to them in any given situation (Gigerenzer and Brighton, 2009). That is, people are not always rationale decision makers. Instead they make use of decision-making heuristics, a form of a mental shortcut, to come to a quick decision based on a limited number of cues. This is known as the cognitive miser approach and contrasts with the naïve scientist approach in which individuals make decisions based on a more comprehensive and thorough evaluation of the information available (Fiske and Taylor, 2013). It has been argued that people are motivated tacticians, in which whether they apply a cognitive miser or naïve scientist approach is in part determined by the urgency, perceived importance and complexity of the situation (Kruglanski, 1996). This strategy reflects limitations in how much information individuals can process at any one time. If we were to attempt to fully process all the information that is available in every situation, we encounter each day, as in the naïve scientist approach, then this would become extremely time consuming (Sweller, 1988). On the other hand, using the cognitive miser approach may be quicker and less cognitively demanding, but is at greater risk of error, as the individual is basing their decision on a limited number of factors. As such individuals switch between strategies based on which they think will be the most optimal for the situation they are facing, an approach which will not always necessarily be correct (Gigerenzer and Brighton, 2009). The tactics used by social engineers are often based on exploiting heuristics, by including elements that encourage the target to engage the cognitive miser approach and make quick, less analytical decisions (Hadnagy, 2018). Examples of this within the social engineering technique of phishing emails can include the use of language that contains emotive elements such as threat, urgency, or financial information (Hadnagy, 2018). However, research connecting information processing to the characteristics of phishing emails is lacking. To fully understand how an individual engages with aspects of a socio-technical system such as phishing emails it is necessary therefore to explore how much and what information they are processing.

One way in which this can be achieved is through eye trackers. Eye tracking technologies are used to measure an individual's eye movements and in turn to determine what they are looking at. This is known as the point of regard and is an indication of where the individual's attention lies. By measuring several factors such as duration of fixations (when the eyes are relatively stationary), the length of saccades (when the eyes move between areas of interest), number of regressions (where the eyes return to a previously fixated point) inferences can be made about much cognitive processing the individual is giving to any part of the stimulus. This information can be combined to explore the scanpath. This refers to the sequence of fixations and eye movements over an image. For example, an eye tracker may be used to determine the scanpath of an individual viewing a web page, which could provide information about the order in which the individual views different parts of the website. This approach has been used extensively in Human-Computer Interaction studies, such as to assess website usability (Cowen et al., 2002). Related technologies can also be used to measure pupil dilation and blink rate, which can measure cognitive overload and fatigue, respectively (Stern et al., 1994; Hossain and Yeasin, 2014). This can be used to help identify possible risk factors, such as if an individual may not be fully processing information being delivered by a complex or sensitive system.

A range of techniques have been used to record eye movements for research since work began in the early 20th century, including methods such as attaching electrodes to the skin around the eye or using contact lenses with an embedded metal coil that can be used to detect eye movements (Poole and Ball, 2006). More recent technologies are less invasive and often involve use of an infra-red camera to infer point-of-regard from the reflection that is given from the cornea, which is the outermost layer of the eye. These cameras can be placed beneath or next to a computer monitor in a way that is unobtrusive. Mobile eye trackers operate using the same principles but are worn in the same manner as a pair of spectacles, which allows for the individual to navigate their environment in a naturalistic style (Cristina and Camilleri, 2018). In the case of cybersecurity this could for instance involve exploring what a social engineer pays attention to when entering the reception area of a company, such as the location of security cameras or the presence of a PC at the reception desk.

This technology has been used to understand user behavior in relation to phishing websites. These are fraudulent websites designed to appear as genuine website, such as for example an internet banking page. Research suggests that only a quarter of people can reliably discriminate between genuine websites and fraudulent websites more than 75% of the time (Iuga et al., 2016). Technological approaches such as spam filters and machine learning may mitigate some of the risk posed by phishing attacks, but it has been argued that technology alone cannot completely prevent this issue (Pfeffel et al., 2019). This highlights the need to better understand the mechanisms behind a successful phishing attack. By using eye tracking it is possible to explore what factors predict whether someone will be tricked by a phishing website, by considering the interaction between the structure of the website and what the person looks at, or indeed fails to look at (Miyamoto et al., 2014). This has been used for example to understand how and if users pay attention to web browser security indicators, such as the Firefox Mozilla SSL certificate (Sobey et al., 2008). Research in this areas has revealed several techniques that have been identified in such phishing websites (Darwish and Bataineh, 2012) each of which can be researched through the use of eye tracking (Miyamoto et al., 2015). This includes the use of similar or related domain names (e.g., replacing a "w" in a website address with a "vv"), the use of high quality of animations to give fraudulent websites a professional feel and the presentation of fake Digital Certificates.

Further uses of eye tracking in cybersecurity have become evident as the research field and technology have continued to develop. For instance it has been demonstrated that the technology can be used to change risky behaviors, such as for example by preventing a user from continuing with use of input forms in a website unless an eye tracker has determined that the individual has looked at the address bar (Miyamoto et al., 2014). Similarly, eye trackers can be used to detect anomalous user behavior. The way in which an individual navigates a system that they are highly familiar with will be different from someone who is less familiar with a system: from work on expertise in visual processing (Miyamoto et al., 2015). Eye movement patterns are highly specialized and detectable when viewing scenes and objects that we are experts at processing (Liversedge et al., 2013). This difference in style will be reflected in eye movements, and could be used as a basis for detecting illicit behavior (Biedert et al., 2012). Recently it has been claimed that eye tracking machines themselves may not be necessary, and that webcams built into phones, laptops and tablets may be sufficient (Krafka et al., 2016) for many of the purposes discussed here. If this is the case, then it removes a major barrier for the adoption of eye tracking related cybersecurity measures in real life situations such as the workplace. As has been noted any technology that is used to protect users from cyber-attack is most effective then it is unobtrusive (Miyamoto et al., 2015).

Whilst there has been research using eye trackers to understand engagement with phishing website there is less research applying this technology to phishing emails (Baki et al., 2017), which are one vector through which targets may be directed to a phishing website in the first instance. There are recommendations made to the public by various organizations around what is likely to denote something as being a phishing email, such as the National Cyber Security Centre advice to look for misspellings, the use of urgency and the use of threatening language (National Cyber Security Centre, 2020), which reflects the typical features of phishing emails identified in the literature (Pfeffel et al., 2019). There is a lack of academic research that has explored the relationship between these features, including how trustworthy such emails are rated and how eye-movements may moderate this relationship. To address this we conducted an exploratory study in which we created phishing emails that employed characteristics and techniques evident in phishing emails, including the presence of misspellings in the sender address, the mention of financial information, the use of threatening language (for example that legal action will be taken if an email is not responded to) and the use of urgency.

MATERIALS AND METHODS

Participants

Twenty-two psychology undergraduates (90% female, age range = 18 to 26, mean age = 20.29) were recruited from a sample responding to an online advertisement. Participants were awarded course credits for their participation.

Design

A within-subjects design was employed in which participants were shown emails that either did or did not include a phishing indicator. There were four types of phishing indicator: financial information, urgency, misspelling, and threat. Stimuli were presented in a random order. Eye-tracking measures used were total dwell time, mean fixation count (denoting interest in a particular content), number of regressions (revisits, indicating that the item required further scrutiny because it drew attention), mean glance duration (denoting depth of processing), entry time and entry sequence (the time and fixation number that an area was attended to, denoting ease of attentional capture).

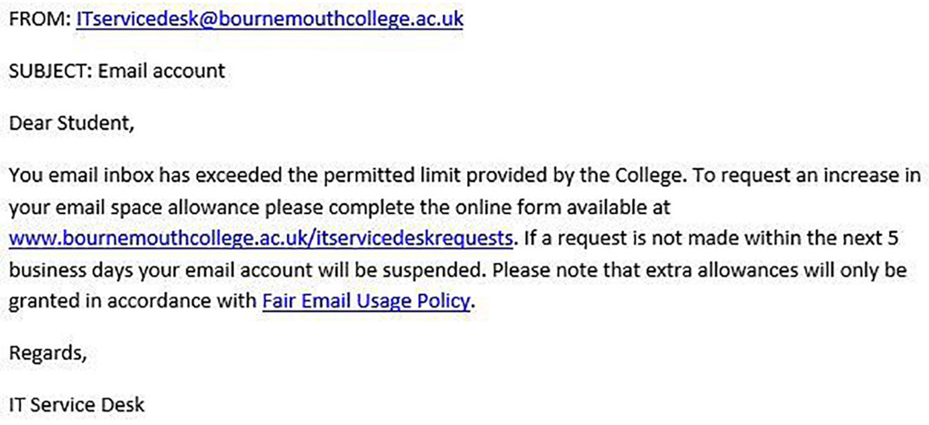
Materials

Thirty-two emails were constructed based upon typical phishing type emails. These were split between the four types of email (misspelling, threatening, urgency, and financial) with four variations of each email type and either containing the phishing email indicator or not. This reflected the elements identified in public guidance from the National Cyber Security Centre on what may be an indicator that an email is a phishing one (National Cyber Security Centre, 2020). These emails were created by the researchers to be relevant to the study sample in terms of names of local organizations and national companies that the email purported to have been sent from. A publicly accessible database of suspected phishing emails[1] was used to guide the creation of the study materials to ensure that these were consistent with phishing emails currently in circulation. The phishing emails created in this study were simple word documents structured according to emails in Microsoft Outlook. These contained a from line with email address, a subject line, the main content with roughly four sentences of text and a by-line.

Areas of interest (AOIs) were mapped onto the emails *post hoc* in BeGaze. This software is used to specify the areas of an image upon which the analysis will be based. These areas of interest were non-overlapping and focused on the core textual information. AOIs were: the email address, subject line, the addressee, the instruction line, any detail (hyperlinks, tracking numbers), and the phishing indicator (financial information, misspelling, threat, and urgency). AOIs were invisible to participants.

Stimuli were presented on an SMI RED 500 eye-tracker with in-built infrared cameras detecting eye movements. The

[1] www.phishtank.com



FROM: ITservicedesk@bournemouthcollege.ac.uk

SUBJECT: Email account

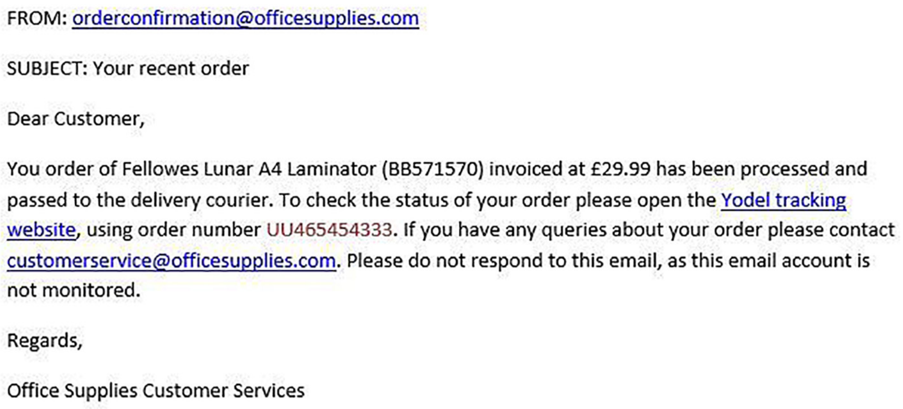
Dear Student,

You email inbox has exceeded the permitted limit provided by the College. To request an increase in your email space allowance please complete the online form available at www.bournemouthcollege.ac.uk/itservicedeskrequests. If a request is not made within the next 5 business days your email account will be suspended. Please note that extra allowances will only be granted in accordance with Fair Email Usage Policy.

Regards,

IT Service Desk

FIGURE 1 | Example of email with urgency indicator.



FROM: orderconfirmation@officesupplies.com

SUBJECT: Your recent order

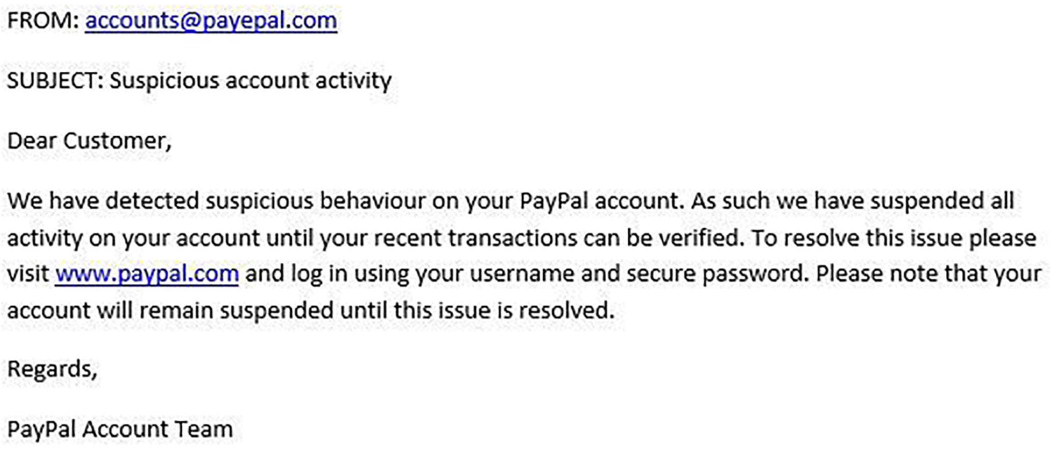
Dear Customer,

You order of Fellowes Lunar A4 Laminator (BB571570) invoiced at £29.99 has been processed and passed to the delivery courier. To check the status of your order please open the Yodel tracking website, using order number UU465454333. If you have any queries about your order please contact customerservice@officesupplies.com. Please do not respond to this email, as this email account is not monitored.

Regards,

Office Supplies Customer Services

FIGURE 2 | Example of email with financial information indicator.



FROM: accounts@payepal.com

SUBJECT: Suspicious account activity

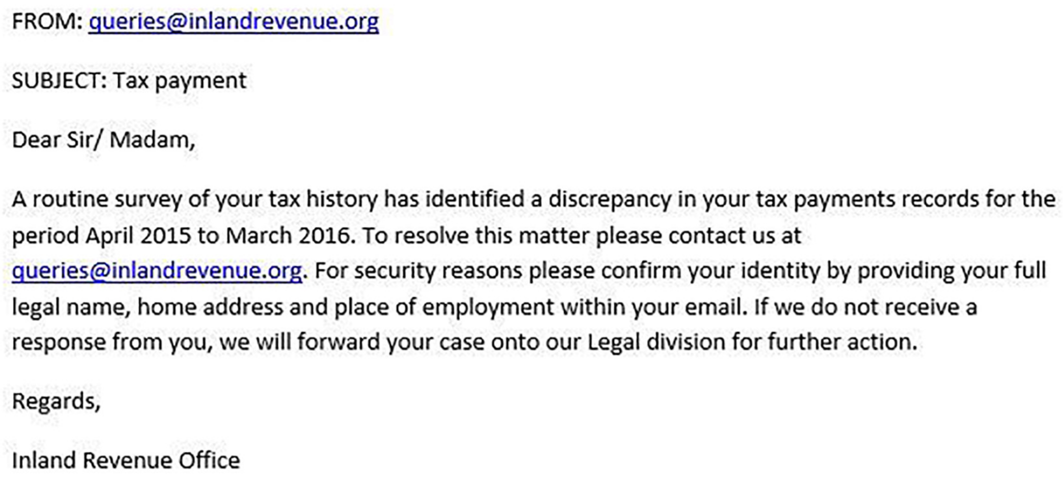
Dear Customer,

We have detected suspicious behaviour on your PayPal account. As such we have suspended all activity on your account until your recent transactions can be verified. To resolve this issue please visit www.paypal.com and log in using your username and secure password. Please note that your account will remain suspended until this issue is resolved.

Regards,

PayPal Account Team

FIGURE 3 | Example of email with misspelling indicator.



FROM: queries@inlandrevenue.org

SUBJECT: Tax payment

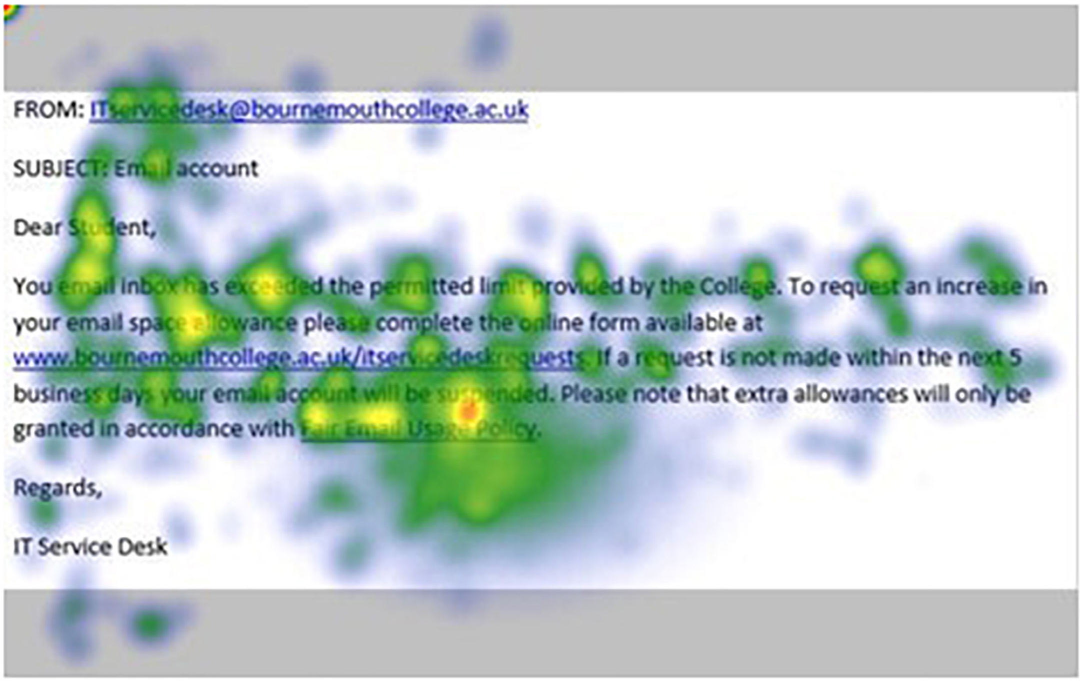
Dear Sir/ Madam,

A routine survey of your tax history has identified a discrepancy in your tax payments records for the period April 2015 to March 2016. To resolve this matter please contact us at queries@inlandrevenue.org. For security reasons please confirm your identity by providing your full legal name, home address and place of employment within your email. If we do not receive a response from you, we will forward your case onto our Legal division for further action.

Regards,

Inland Revenue Office

FIGURE 4 | Example of email with threat indicator.



FROM: ITservicedesk@bournemouthcollege.ac.uk

SUBJECT: Email account

Dear Student,

You email inbox has exceeded the permitted limit provided by the College. To request an increase in your email space allowance please complete the online form available at www.bournemouthcollege.ac.uk/itservicedeskrequests. If a request is not made within the next 5 business days your email account will be suspended. Please note that extra allowances will only be granted in accordance with Fair Email Usage Policy.

Regards,

IT Service Desk

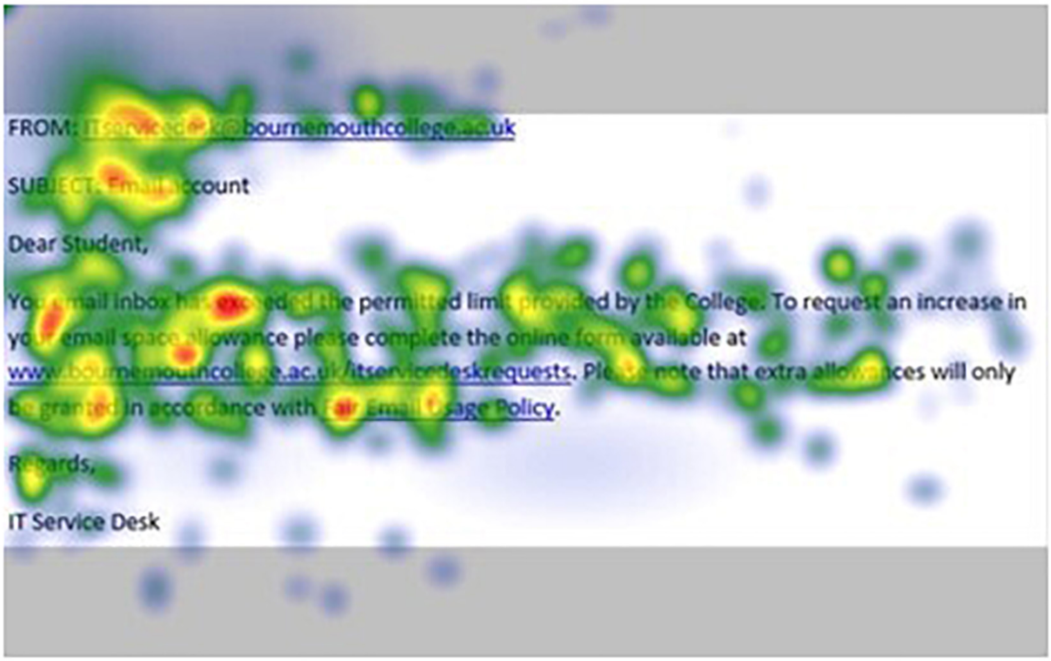
FIGURE 5 | Heat maps averaged across all participants for email with urgency indicator.

screen was a 22-inch high-resolution LCD. Eye movements were recorded at 500 Hz with an accuracy of 0.4° of visual angle using SMI iView.

Procedure

Piloting was conducted with a sample of 8 postgraduate research students. The purpose of this was to test the feasibility of using the eye tracker facilities for the intended purposes of the study. These trials involved participants using the same equipment to view examples of phishing emails. These emails were not split by type and participants were not asked to provide any rating of the emails. No technological or methodological problems were identified during this piloting phase.

Once piloting was completed the main study commenced. After providing informed consent, participants were told that they would be viewing a series of emails and that they would be asked to give a rating of how trustworthy they felt each email appeared to be. Participants' eyes were then calibrated to the eye tracker using the standard in-built 9-point calibration procedure. Following calibration, the eye tracking



FROM: ITservicedesk@bournemouthcollege.ac.uk

SUBJECT: Email account

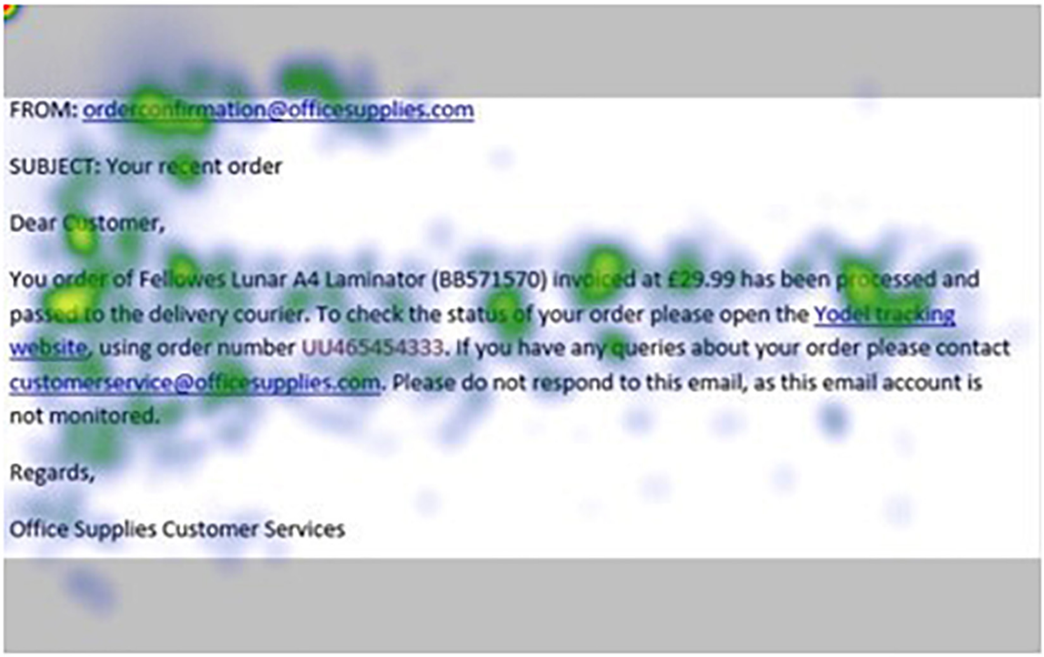
Dear Student,

Your email inbox has exceeded the permitted limit provided by the College. To request an increase in your email space allowance please complete the online form available at www.bournemouthcollege.ac.uk/itservicedeskrequests. Please note that extra allowances will only be granted in accordance with Fair Email Usage Policy.

Regards,

IT Service Desk

FIGURE 6 | Heat maps averaged across all participants for matched email without urgency indicator.



FROM: orderconfirmation@officesupplies.com

SUBJECT: Your recent order

Dear Customer,

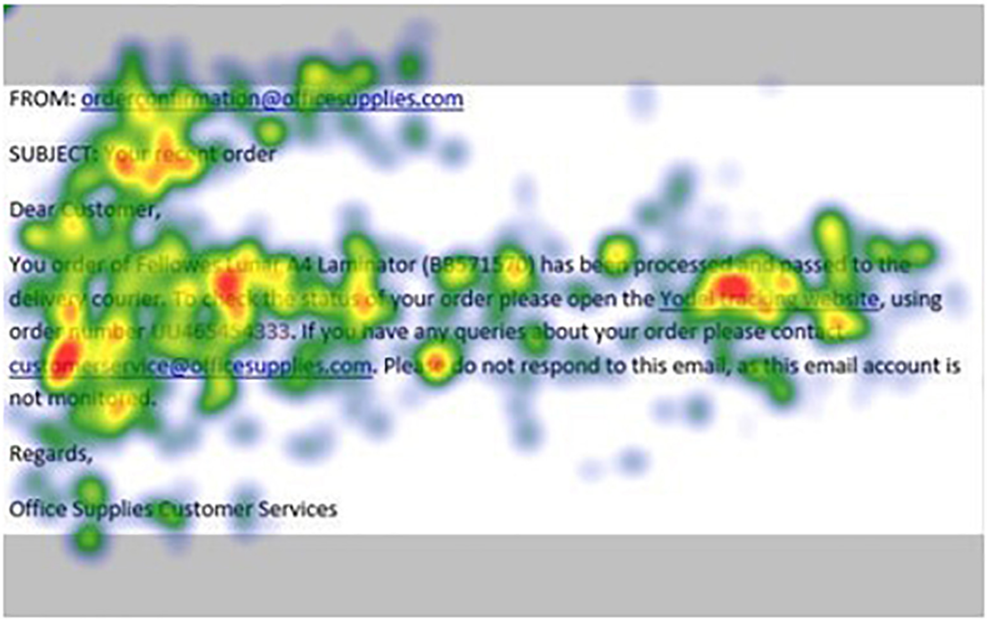
You order of Fellowes Lunar A4 Laminator (BB571570) invoiced at £29.99 has been processed and passed to the delivery courier. To check the status of your order please open the Yodel tracking website, using order number UU465454333. If you have any queries about your order please contact customerservice@officesupplies.com. Please do not respond to this email, as this email account is not monitored.

Regards,

Office Supplies Customer Services

FIGURE 7 | Heat maps averaged across all participants for email with financial information indicator.

was validated, to ensure consistent and accurate tracking. Validation consisted of the standard SMI calibration and validation procedure. Participants were requested to follow a ball around to 7-pseudo random locations around the screen. Calibration was considered successful if the eyes were calibrated within 1° of visual angle. If calibration failed, the participant was recalibrated once, otherwise they were removed from the analysis. The calibration was validated using the default procedure - participants eyes fixated on the center of the screen and if this was recording accurately, the trial proceeded.



FROM: orderconfirmation@officesupplies.com

SUBJECT: Your recent order

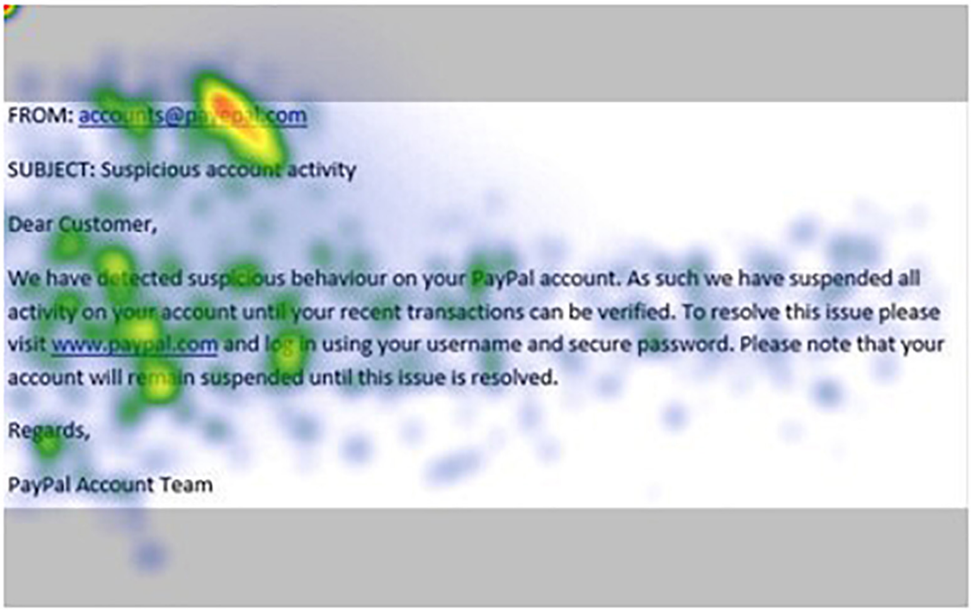
Dear Customer,

You order of Fellowes Lunar A4 Laminator (B8571570) has been processed and passed to the delivery courier. To check the status of your order please open the Yodel tracking website, using order number UU465454333. If you have any queries about your order please contact customerservice@officesupplies.com. Please do not respond to this email, as this email account is not monitored.

Regards,

Office Supplies Customer Services

FIGURE 8 | Heat maps averaged across all participants for matched email without financial information indicator.



FROM: accounts@[illegible].com

SUBJECT: Suspicious account activity

Dear Customer,

We have detected suspicious behaviour on your PayPal account. As such we have suspended all activity on your account until your recent transactions can be verified. To resolve this issue please visit www.paypal.com and log in using your username and secure password. Please note that your account will remain suspended until this issue is resolved.

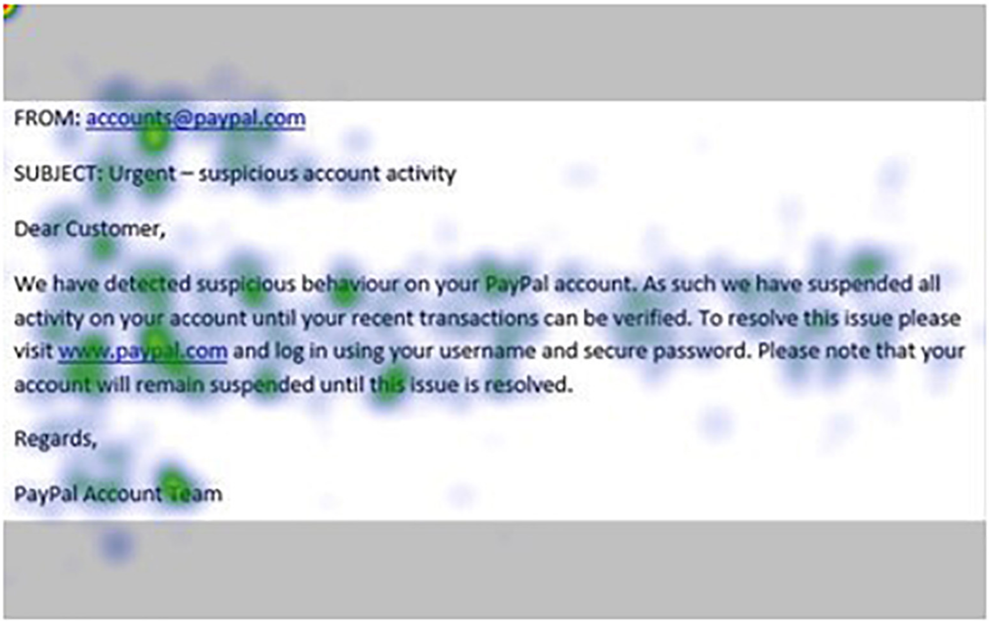
Regards,

PayPal Account Team

FIGURE 9 | Heat maps averaged across all participants for email with misspelling indicator.

This validation was repeated after every 13 trials. Following this, participants began the experimental task. There were 32 identical trials.

In each trial, participants saw a blank fixation screen lasting 500 ms. Following this, participants saw the email. For each, participants were tasked with reading the email ready to rate it for trustability. Each email was on screen for 10 s. This time was chosen to represent the rather short amount of time that is devoted to reading each email that individuals receive (Hart, 2017). After the email, participants were given the rating screen, in which they were visually asked to rate how trustworthy the preceding email was on an 8-point Likert-type scale with the anchor points "Not at all trustworthy" and "Highly trustworthy." Participants notified the researcher verbally of their choice, who then entered their answer into a numerical keypad. This was done to avoid unnecessary head movements by the



FROM: accounts@paypal.com

SUBJECT: Urgent – suspicious account activity

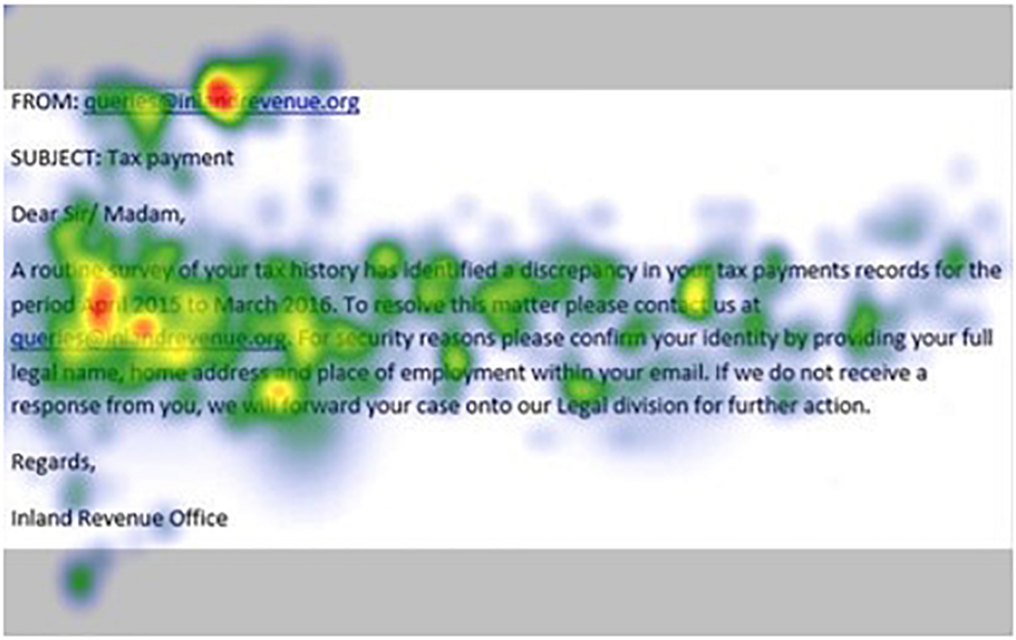
Dear Customer,

We have detected suspicious behaviour on your PayPal account. As such we have suspended all activity on your account until your recent transactions can be verified. To resolve this issue please visit www.paypal.com and log in using your username and secure password. Please note that your account will remain suspended until this issue is resolved.

Regards,

PayPal Account Team

FIGURE 10 | Heat maps averaged across all participants for matched email without misspelling indicator.



FROM: queries@inlandrevenue.org

SUBJECT: Tax payment

Dear Sir/ Madam,

A routine survey of your tax history has identified a discrepancy in your tax payments records for the period April 2015 to March 2016. To resolve this matter please contact us at queries@inlandrevenue.org. For security reasons please confirm your identity by providing your full legal name, home address and place of employment within your email. If we do not receive a response from you, we will forward your case onto our Legal division for further action.

Regards,

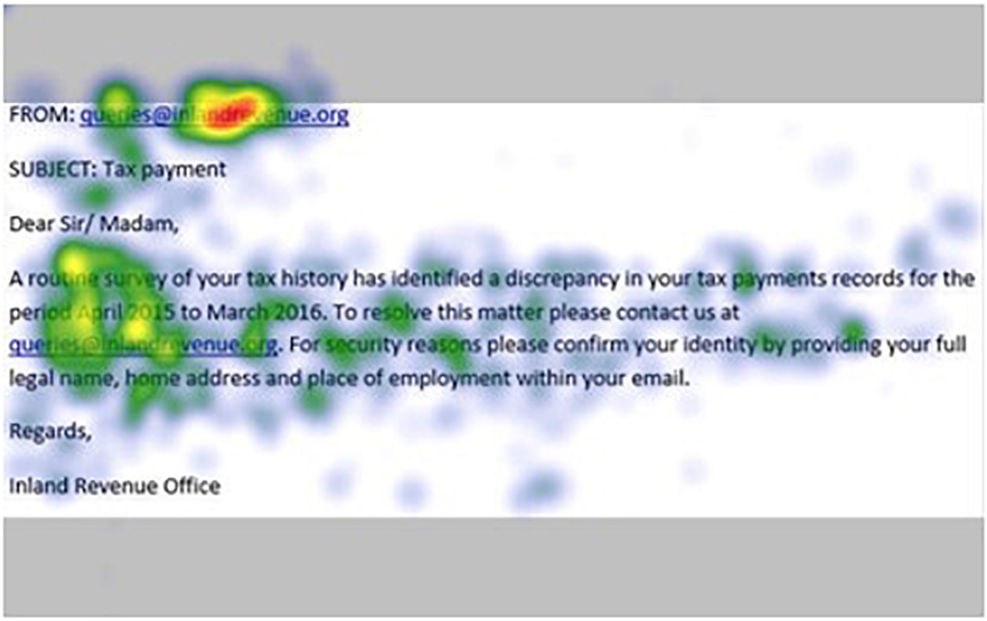
Inland Revenue Office

FIGURE 11 | Heat maps averaged across all participants for email with threat indicator.

participant. Following completion of all trials, participants were thanked and debriefed.

Analysis Protocol

We assessed first whether the AOI containing the phishing indicator was scanned. To assess this, we analyzed whether the phishing indicator AOI was scanned more than would be expected by chance. For this, we ran a series of Bonferroni-corrected (α = 0.0125) one-subjects *t*-tests (two-tailed) comparing to a chance value for the region (which was based on the AOI size relative to the size of the screen). Secondly, we analyzed the amount of scanning to the other AOIs with and without the phishing indicator. Because the AOIs filled proportionally less of the screen in the phishing



FROM: queries@inlandrevenue.org

SUBJECT: Tax payment

Dear Sir/ Madam,

A routine survey of your tax history has identified a discrepancy in your tax payments records for the period April 2015 to March 2016. To resolve this matter please contact us at queries@inlandrevenue.org. For security reasons please confirm your identity by providing your full legal name, home address and place of employment within your email.

Regards,

Inland Revenue Office

FIGURE 12 | Heat maps averaged across all participants for matched email without threat indicator.

indicator present conditions, we area-normalized the AOIs by calculating the proportion of the screen that the AOIs occupied in each screen.

Our secondary analysis concerned which type of phishing indicator was most detectable. This was assessed with a one-way-ANOVA on the area-normalized phishing AOIs. For all analyses, the assumptions of parametric data were tested: Whenever Mauchley's test of sphericity was significant, the Huynh-Feldt correction was applied to the degrees of freedom. If tests of normality were violated, a non-parametric test was used. For *post hoc* tests, the *p*-values were adjusted for multiple comparisons.

RESULTS

Figures 1–4 show examples of emails with the four types of phishing indicator. **Figures 5–12** show a series of heat maps indicating where participants scanned images, split into the four pairs of emails either with or without the phishing email indicator (financial information, misspelling, threat, and urgency). We present an example of each category of phishing email with and without the phishing content for ease of understanding.

Our analysis protocol was applied, and summary statistics for the one-sample *t*-tests are shown in **Table 1**. Specifically, these results show that while the phishing AOIs were scanned (denoted by fixation count) and revisited (regression count) more frequently with more intense scanning (glance duration) than one would expect by chance, the total duration of scanning (dwell time) was less then would be expected by chance. In other words, less time was spent viewing the phishing indicators even though they required greater attentional resources paid to them.

TABLE 1 | Mean (and standard error of the mean) for total dwell time (area-normalized ms), mean fixation count, number of regressions, and mean glance duration (ms), with one-sample *t*-value (df), and significance level.

		Mean	***t*-value**	***p*-value**
Financial information	Dwell time (*ms*)	159 (23)	−9.06 (20)	<0.001
	Fixation count	1.40 (0.40)	13.64 (14)	<0.001
	Regression count	0.22 (0.09)	2.51 (14)	=0.025
	Glance Duration (*ms*)	93 (12)	7.87 (19)	<0.001
Misspelling	Dwell time (*ms*)	479 (101)	−3.03 (20)	=0.007
	Fixation count	3.31 (0.63)	5.25 (17)	<0.001
	Regression count	1.55 (0.44)	3.50 (17)	=0.003
	Glance Duration (*ms*)	163 (20)	8.33 (20)	<0.001
Threatening content	Dwell time (*ms*)	629 (50)	−74.48 (20)	<0.001
	Fixation count	3.18 (0.90)	15.83 (19)	<0.001
	Regression count	1.47 (0.19)	7.79 (19)	<0.001
	Glance Duration (*ms*)	104 (37)	12.91 (20)	<0.001
Urgency content	Dwell time (*ms*)	709 (78)	−49.68 (20)	<0.001
	Fixation count	3.50 (0.45)	7.72 (19)	<0.001
	Regression count	1.46 (0.31)	4.78 (19)	<0.001
	Glance Duration (*ms*)	104 (44)	10.94 (20)	<0.001

Degrees of freedom are lower due to missing values for some participants.

Our second analysis focused on exploring whether the presence of the phishing indicator affected the scanning of the other content. The presence of each type of phishing indicator did not significantly affect normalized dwell time, $F(1,20) = 0.06$, $MSE = 28486$, $p = 0.813$, $\eta_p^2 < 0.01$. **Figure 13** shows the mean dwell duration to each of

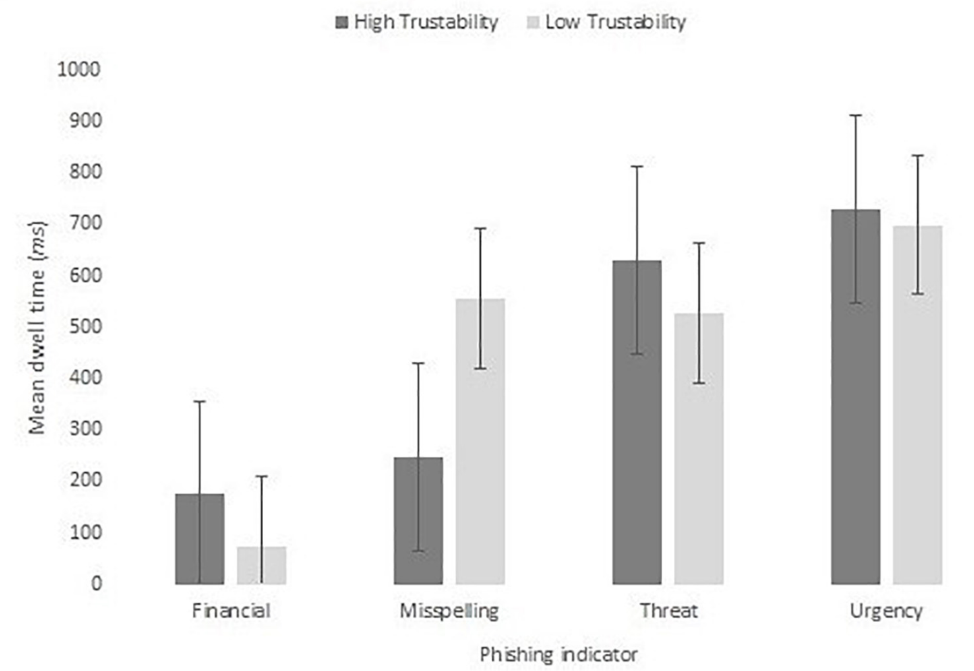


FIGURE 13 | Mean dwell time to the scam item for high and low trustability emails split by indicator type.

the area-normalized AOIs for those with and without the phishing content.

Our final analysis concerned which type of phishing indicator would be more noticeable. **Table 2** shows the means for each eye-tracking measure. Phishing indicator type affected: total dwell time, $F(3,39) = 4.98$, $MSE = 800312348$, $p = 0.031$, $\eta_p^2 = 0.28$, fixation count, $F(3,39) = 6.30$, $MSE = 20.29$, $p = 0.014$, $\eta_p^2 = 0.33$, regression count, $F(3,39) = 6.72$, $MSE = 0.95$, $p = 0.003$, $\eta_p^2 = 0.34$, glance duration, $F(3,57) = 5.89$, $MSE = 68.76$, $p = 0.004$, $\eta_p^2 = 0.24$, entry time, $F(3,39) = 8.24$, $MSE = 0.5184111$, $p = 0.003$, $\eta_p^2 = 0.34$, and sequence, $F(3,39) = 4.72$, $MSE = 1.91$, $p = 0.024$, $\eta_p^2 = 0.27$.

Specifically, financial indicators were viewed for less time than threat indicators (mean difference = 469, $p = 0.015$, $r = 0.79$) and urgency indicators (mean difference = 550, $p = 0.019$, $r = 0.72$). Further, they were viewed less frequently with less regressions than threatening indicators (mean $\text{difference}_{\text{fixationcount}} = 2.09$, $p < 0.001$, $r = 0.61$, mean $\text{difference}_{\text{regressioncount}} = 1.36$, $p < 0.001$, $r = 0.97$) and urgency indicators (mean $\text{difference}_{\text{fixationcount}} = 1.92$, $p < 0.001$, $r = 0.89$, mean $\text{difference}_{\text{regressioncount}} = 0.92$, $p < 0.001$, $r = 0.59$). Glance duration was shorter for threat indicators than financial indicators (mean difference = 60.74, $p = 0.044$, $r = 0.40$). Threat and urgency indicators were viewed earlier than misspelling indicators (threat: mean $\text{difference}_{\text{entrytime}} = 1706$ $p < 0.001$, $r = 0.56$, mean $\text{difference}_{\text{sequence}} = 0.93$, $p < 0.001$, $r = 0.33$; urgency: mean $\text{difference}_{\text{entry}}$ time = 2855, $p < 0.001$, $r = 0.74$ and mean $\text{difference}_{\text{sequence}} = 1.33$, $p = 0.011$, $r = 0.52$).

TABLE 2 | Mean (and standard error of the mean) for total dwell time (area-normalized ms), mean fixation count, number of regressions, mean glance duration (ms), entry time (ms), and sequence.

	Financial phishing indicator	Misspelling phishing indicator	Threatening phishing indicator	Urgency phishing indicator
Total dwell time (ms)	159 (23)	479 (101)	630 (50)	709 (78)
Fixation count	1.43 (0.11)	2.59 (0.44)	3.35 (0.15)	3.52 (0.51)
regressions count	0.24 (0.09)	0.92 (0.24)	1.60 (0.23)	1.22 (0.24)
Glance duration (ms)	93 (12)	165 (20)	104 (9)	105 (10)
Entry time (ms)	3712 (529)	4882 (1474)	1603 (241)	1395 (247)
Sequence	3.52 (0.28)	3.23 (0.36)	4.16 (0.34)	4.56 (0.21)

A further set of analyses were run on the trustability ratings, shown in **Figure 14**. These were subjected to a 2 × 4 within-subjects ANOVA. This revealed a main effect of phishing indicator, $F(3,60) = 25.63$, $MSE = 1.50$, $p < 0.001$, $\eta_p^2 = 0.58$. Emails with misspelling and threatening phishing indicators were rated as less trustworthy than financial (mean difference = 2.02, $p < 0.001$, $r = 0.76$, mean difference = 1.75, $p < 0.001$, $r = 0.70$) and urgency (mean difference = 1.44, $p < 0.001$, $r = 0.62$, mean difference = 1.16, $p = 0.002$, $r = 0.53$) scams. There was also a main effect of presence of phishing indicator, $F(1,20) = 10.87$, $MSE = 0.74$, $p = 0.004$, $\eta_p^2 = 0.35$, in which phishing indicators present emails (5.27, $SE = 0.21$) were rated as less trustworthy than emails without phishing indicators (4.83, $SE = 0.16$). However, these effects interacted, $F(3,60) = 9.45$, $MSE = 0.97$, $p < 0.001$, $\eta_p^2 = 0.32$. The interaction was revealed by the effect of phishing indicator presence only being

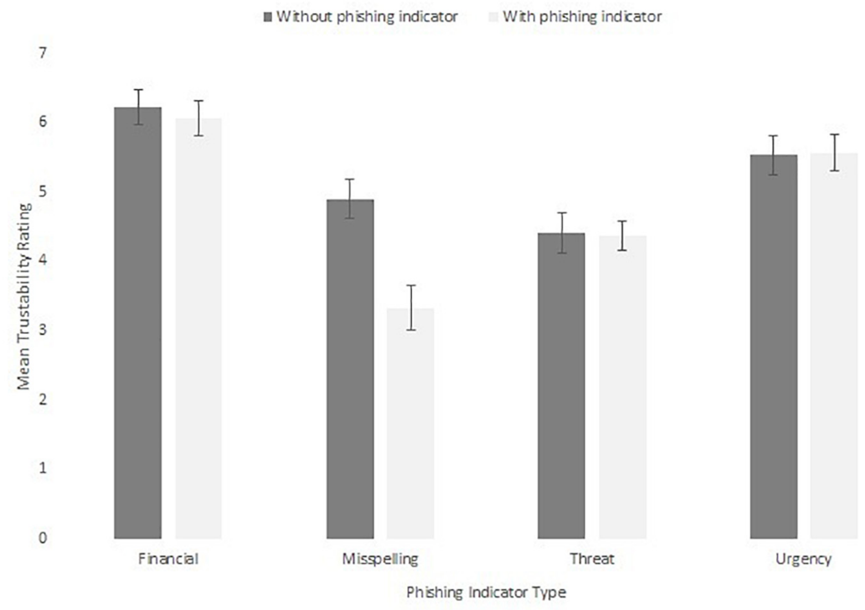


FIGURE 14 | Mean trustability ratings for each phishing indicator.

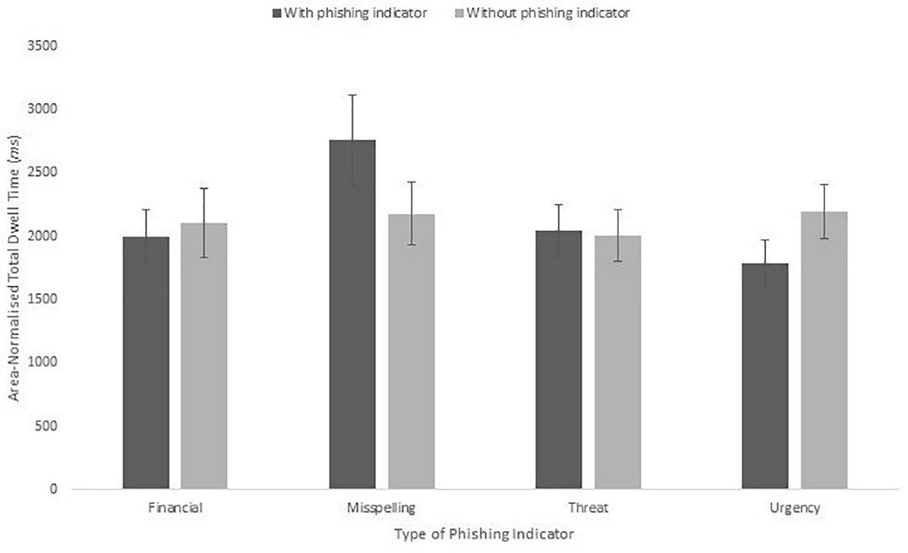


FIGURE 15 | Mean area-normalized total dwell time split by indicator presence and absence.

significant for the misspelling scam, $t(20) = 4.05$, $p = 0.001$, $r = 0.50$, and not for the other types of scams (smallest $p = 0.397$).

Finally, we assessed whether the trustability rating influenced the amount of scanning to the phishing indicators of the emails. We used two protocols to assess this. In the first we ran a series of correlations between the dwell time for each email type and the trustability rating given. None of these correlations were significant: financial phishing indicators, $r(19) = 0.20$, $p = 0.385$; misspelling, $r(19) = -0.41$, $p = 0.063$; threat, $r(19) = 0.10$, $p = 0.659$; and urgency, $r(19) = -0.26$, $p = 0.263$. In the second, we analyzed whether dwell time to the phishing indicator item was different for emails rated as trustable (scoring higher than 4) compared to those rated as untrustable (rated 4 or lower) split by type of phishing indicator, shown in **Figure 15**. This analysis was done by-item. The resulting 4 × 2 mixed ANOVA showed no significant effect of trustability, $F(1,12) = 0.06$, $MSE = 45904$, $p = 0.811$, $\eta_p^2 = 0.01$, nor an interaction with phishing indicator type, $F(1,12) = 1.68$ $MSE = 45904$, $p = 0.223$, $\eta_p^2 = 0.30$.

DISCUSSION

The results of the study were notable in several ways. Participants spent less time overall looking at indicators of phishing than they would be expected to by chance. In addition, the presence of phishing indicators did not significantly impact on how much time is spent looking at the rest of the email. Overall, this may suggest that individuals require little processing time to recognize elements that relate to phishing. The phishing variants of each email were also rated as being less trustworthy than the non-phishing variants, suggesting that participants have some ability to recognize that the selected features are associated with fraudulent emails. Yet there was no statistically significant association between the trustworthiness rating and the total scanning time for the phishing indicators within the emails. As such whilst emails with phishing indicators were rated as less trustworthy than those without, this does not appear to be explained by how much time is spent attending to those phishing indicators. This makes it unclear whether the features of phishing emails that would appear to be designed to capture attention, exploit heuristics and invoke a cognitive miser style of processing are achieving this. An interpretation of this could be that the relationship between the presence of features related to phishing emails and how trustworthy that email is seen to be is more complex than expected. Similar unexpected, complex and inconsistent results have been found in relation to susceptibility to phishing emails and other factors including personality, knowledge of computers and gender (Kleitman et al., 2018).

Other aspects of the results were more in keeping with previous research. For instance, it was noted that participants would tend to look first at phishing indicators relating to urgency and threats before looking at misspellings and financial information. This could be a reflection of survival information bias (Nairne, 2010), in which individuals place priority on processing information that may relates to their well-being. Emails containing misspelling were also rated as being less trustworthy than the other emails, which may be due to the presence of misspelling being a more categorical factor than the use of urgency or threatening language, which are open to interpretation. Financial phishing email indicators were associated with the least frequent number of fixations and the least amount of overall dwell time, as compared to the phishing indicators in the misspelling, urgency, and threat email variations. Emails with financial phishing indicators were also rated as being more trustworthy than emails with misspelling or threat phishing indicators. This suggests that the inclusion of financial information within phishing emails has a lower impact of how that email is processed and to what degree it is trusted.

There were limitations to this study. A relatively small sample size was used, although this is not atypical when compared to other eye-tracking studies (Tecce et al., 1998; Libben and Titone, 2009; Choi et al., 2017). While the sample size was consistent with previous eye-tracking research, it is not sufficient to explore individual variability in how well eye movements predict ability to spot phishing emails. Further recruitment of participants was not possible due to constraints caused by the COVID-19 situation.

The participants consisted of a narrow demographic from a single geographical location. The sample was also predominantly female. There is no evidence of gender differences in eye movements (Klein and Ettinger, 2019) and a lack of consistent research on the role of gender in phishing email susceptibility (Kleitman et al., 2018). Nevertheless, having a more diverse sample may help identify if there are certain types of phishing email that are more impactful on different demographic groups. Due to the limited research in this area there was also a lack of baseline evidence to use to inform the creation of phishing email materials. Examples of phishing emails available on websites such as www.phishtank.com are not ideally suited to experimental designs, as they often include conflation of different phishing techniques, such as a combination of threat and urgency. We opted to create our own stimuli in this study to reduce the influence of such possible confounders, however, it is difficult to do so completely whilst keeping the stimuli realistic. Further refinement of these stimuli may also help clarify the relationship between content and how phishing emails are read and judged. Finally, we note that asking participants to provide a trustworthiness rating of the stimuli may have alerted them that the study related to phishing emails. As demonstrated by Parsons et al. (2015) participants may be more successful at identifying phishing emails when they are aware in advance that they may be about to do so.

The results of this study demonstrate some important points. It provides evidence that eye tracking technology can be used to determine whether people look at the common indicators of phishing emails, and also inform us on the order in which these are attended to. In doing so it also demonstrated some unexpected patterns, including that individuals look at phishing indicators more frequently than would be expected by chance but, counterintuitively, spend less overall time doing so than would be expected by chance. Building upon this research may provide more avenues for the understanding and mitigation of the serious threat that phishing emails pose to cybersecurity.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Science, Technology & Health Research Ethics Panel, Bournemouth University. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

JM contributed to the content relating to phishing emails and social engineering. PH contributed to the content on eye-tracking technology and led the analysis. All authors contributed to the article and approved the submitted version.

REFERENCES

Baki, S., Verma, R., Mukherjee, A., and Gnawali, O. (2017). "Scaling and effectiveness of email masquerade attacks: exploiting natural language generation," in *Paper Presented at the Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*, Abu Dhabi.

Biedert, R., Frank, M., Martinovic, I., and Song, D. (2012). "Stimuli for Gaze Based Intrusion Detection," in *Future Information Technology, Application, and Service: FutureTech 2012*, Vol. 1, eds J. J. Park, V. C. M. Leung, C.-L. Wang, and T. Shon (Dordrecht: Springer Netherlands), 757–763. doi: 10.1007/978-94-007-4516-2_80

Choi, W., Lowder, M. W., Ferreira, F., Swaab, T. Y., and Henderson, J. M. (2017). Effects of word predictability and preview lexicality on eye movements during reading: a comparison between young and older adults. *Psychol. Aging* 32, 232–242. doi: 10.1037/pag0000160

Cowen, L., Ball, L. J. S., and Delin, J. (2002). "An Eye Movement Analysis of Web Page Usability," in *Paper Presented at the People and Computers XVI - Memorable Yet Invisible: Proceedings of HCI* 2002, London.

Cristina, S., and Camilleri, K. P. (2018). Unobtrusive and pervasive video-based eye-gaze tracking. *Image Vis. Comput.* 74, 21–40. doi: 10.1016/j.imavis.2018.04.002

Darwish, A., and Bataineh, E. (2012). "Eye tracking analysis of browser security indicators," in *Paper Presented at the 2012 International Conference on Computer Systems and Industrial Informatics*, Sharjah.

Das, A., Baki, S., Aassal, A. E., Verma, R. M., and Dunbar, A. (2020). SoK: a comprehensive reexamination of phishing research from the security perspective. *IEEE Commun. Surv. Tutor.* 22, 671–708. doi: 10.1109/comst.2019.2957750

Fiske, S. T., and Taylor, S. E. (2013). *Social Cognition: From Brains to Culture*, 2nd Edn. Nwe York, NY: Sage Publishing.

Gigerenzer, G., and Brighton, H. (2009). Homo heuristicus: why biased minds make better inferences. *Top. Cogn. Sci.* 1, 107–143. doi: 10.1111/j.1756-8765.2008.01006.x

Hadnagy, C. (2018). *Social Engineering: The Science of Human Hacking*. Indianapolis, IN: Wiley.

Hart, E. (2017). *Average Email Reading Time Increase*. Avaliable at: https://blog.dotdigital.com/average-email-reading-time-increases/ (accessed June 15, 2020).

Hossain, G., and Yeasin, M. (2014). "Understanding effects of cognitive load from pupillary responses using hilbert analytic phase," in *Paper presented at the 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, Columbus, OH.

Iuga, C., Nurse, J. R. C., and Erola, A. (2016). Baiting the hook: factors impacting susceptibility to phishing attacks. *Hum. Centric Comput. Inform. Sci.* 6:8. doi: 10.1186/s13673-016-0065-2

Klein, C., and Ettinger, U. (eds) (2019). *Eye Movement Research: An Introduction to Its Scientific Foundations and Applications*. Berlin: Springer International Publishing.

Kleitman, S., Law, M. K. H., and Kay, J. (2018). It's the deceiver and the receiver: individual differences in phishing susceptibility and false positives with item profiling. *PLoS One* 13:e0205089. doi: 10.1371/journal.pone.0205089

Krafka, K., Khosla, A., Kellnhofer, P., Kannan, H., Bhandarkar, S., Matusik, W., et al. (2016). "Eye tracking for everyone," in *Paper Presented at the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV.

Kruglanski, A. W. (1996). "Motivated social cognition: principles of the interface," in *Social Psychology: Handbook of Basic Principles*, eds E. T. Higgins and A. W. Kruglanski (New York, NY: Guilford Press).

Libben, M. R., and Titone, D. A. (2009). Bilingual lexical access in context: evidence from eye movements during reading. *J. Exp. Psychol. Learn. Mem. Cogn.* 35, 381–390. doi: 10.1037/a0014875

Liversedge, S. P., Gilchrist, I. D., and Everling, S. (2013). *The Oxford Handbook of Eye Movements*. Oxford: Oxford University Press.

McCalley, H., Wardman, B., and Warner, G. (2011). "Analysis of back-doored phishing kits," in *Paper Presented at the IFIP International Conference on Digital Forensics*, Berlin.

Miyamoto, D., Blanc, G., and Kadobayashi, Y. (2015). "Eye can tell: on the correlation between eye movement and phishing identification," in *Paper Presented at the Neural Information Processing*, Cham.

Miyamoto, D., Iimura, T., Blanc, G., Tazaki, H., and Kadobayashi, Y. (2014). "EyeBit: eye-tracking approach for enforcing phishing prevention habits," in *Paper Presented at the 2014 Third International Workshop on Building Analysis Datasets and Gathering Experience Returns for Security (BADGERS)*, Wrocław.

Nairne, J. S. (2010). "Adaptive memory: evolutionary constraints on remembering," in *Psychology of Learning and Motivation*, Vol. 53, ed. H. R. Brian (Cambridge: MA: Academic Press), 1–32.

National Cyber Security Centre (2020). *I've Received a Suspicious Email: Our Guide to Spotting and Dealing With Phishing Emails*. Avaliable at: https://www.ncsc.gov.uk/guidance/suspicious-email-actions (accessed June 15, 2020).

Parsons, K., McCormac, A., Pattinson, M., Butavicius, M., and Jerram, C. (2015). The design of phishing studies: challenges for researchers. *Comput. Secur.* 52, 194–206. doi: 10.1016/j.cose.2015.02.008

Pfeffel, K., Ulsamer, P., and Müller, N. H. (2019). "Where the user does look when reading phishing mails – an eye-tracking study," in *Learning and Collaboration Technologies. Designing Learning Experiences. HCII 2019. Lecture Notes in Computer Science*, eds P. Zaphiris and A. Ioannou (Cham: Springer International Publishing), 277–287. doi: 10.1007/978-3-030-21814-0_21

Poole, A., and Ball, L. J. (2006). "Eye tracking in human-computer interaction and usability research: current status and future prospects," in *The Encyclopedia of Human-Computer Interaction*, ed. Interaction Design Foundation (Calgary, AB: Idea Group Inc).

Salahdine, F., and Kaabouch, N. (2019). Social engineering attacks: a survey. *Future Internet* 11:17. doi: 10.3390/fi11040089

Sobey, J., Biddle, R., van Oorschot, P. C., and Patrick, A. S. (2008). "Exploring user reactions to new browser cues for extended validation certificates," in *Computer Security - ESORICS 2008: 13th European Symposium on Research in Computer Security, Málaga, Spain, October 6-8, 2008. Proceedings*, eds S. Jajodia and J. Lopez (Berlin: Springer Berlin Heidelberg), 411–427. doi: 10.1007/978-3-540-88313-5_27

Stern, J. A., Boyer, D., and Schroeder, D. (1994). Blink rate: a possible measure of fatigue. *Hum. Fact.* 36, 285–297. doi: 10.1177/001872089403600209

Sweller, J. (1988). Cognitive load during problem solving: effects on learning. *Cogn. Sci.* 12, 257–285. doi: 10.1207/s15516709cog1202_4

Tecce, J. J., Gips, J., Olivieri, C. P., Pok, L. J., and Consiglio, M. R. (1998). Eye movement control of computer functions. *Int. J. Psychophysiol.* 29, 319–325. doi: 10.1016/s0167-8760(98)00020-8

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 25 September 2020
doi: 10.3389/fpsyg.2020.535803

Influence of Network Size on Adversarial Decisions in a Deception Game Involving Honeypots

Harsh Katakwar[1], *Palvi Aggarwal*[2], *Zahid Maqbool*[1] *and Varun Dutt*[1]*

[1] Applied Cognitive Science Laboratory, Indian Institute of Technology Mandi, Kamand, India, [2] Dynamic Decision Making Laboratory, Carnegie Mellon University, Pittsburgh, PA, United States

Deception via honeypots, computers that pretend to be real, may provide effective ways of countering cyberattacks in computer networks. Although prior research has investigated the effectiveness of timing and amount of deception via deception-based games, it is unclear as to how the size of the network (i.e., the number of computer systems in the network) influences adversarial decisions. In this research, using a deception game (DG), we evaluate the influence of network size on adversary's cyberattack decisions. The DG has two sequential stages, probe and attack, and it is defined as DG (n, k, γ), where n is the number of servers, k is the number of honeypots, and γ is the number of probes that the adversary makes before attacking the network. In the probe stage, participants may probe a few web servers or may not probe the network. In the attack stage, participants may attack any one of the web servers or decide not to attack the network. In a laboratory experiment, participants were randomly assigned to a repeated DG across three different between-subject conditions: small (20 participants), medium (20 participants), and large (20 participants). The small, medium, and large conditions used DG (2, 1, 1), DG (6, 3, 3), and DG (12, 6, 6) games, respectively (thus, the proportion of honeypots was kept constant at 50% in all three conditions). Results revealed that in the small network, the proportions of honeypot and no-attack actions were 0.20 and 0.52, whereas in the medium (large) network, the proportions of honeypot and no-attack actions were 0.50 (0.50) and 0.06 (0.03), respectively. There was also an effect of probing actions on attack actions across all three network sizes. We highlight the implications of our results for networks of different sizes involving deception via honeypots.

Keywords: honeypot, cybersecurity, cyber deception, deception game, adversary, defender, probes, attacks

OPEN ACCESS

Edited by:
Richard Keith Wortley,
University College London,
United Kingdom

Reviewed by:
Stefan Sütterlin,
Østfold University College, Norway
Yilun Shang,
Northumbria University,
United Kingdom

***Correspondence:**
Varun Dutt
varun@iitmandi.ac.in

Specialty section:
This article was submitted to
Cognition,
a section of the journal
Frontiers in Psychology

Received: *17 February 2020*
Accepted: *24 August 2020*
Published: *25 September 2020*

Citation:
Katakwar H, Aggarwal P,
Maqbool Z and Dutt V (2020)
Influence of Network Size on
Adversarial Decisions in a Deception
Game Involving Honeypots.
Front. Psychol. 11:535803.
doi: 10.3389/fpsyg.2020.535803

INTRODUCTION

Cyberattacks, organized attempts to disable computers, steal data, or compromise websites, have been steadily increasing (Trustwave, 2019). For example, there was a rise of 56% in detected web-based cyberattacks on enterprise networks in 2018 compared to 2017 (Symantec, 2019). Some of the detected web-based attacks in 2018 included SQL injection, path traversal, and cross-site scripting, which accounted for more than 50% of cyberattacks on corporate resources (PosTech, 2020).

Due to the prevalence of different kinds of cyberattacks and the associated cyber-defense costs (Hope, 2020), one may need to develop and evaluate technologies that provide security against cyberattacks (Sayegh, 2020). Currently, there are a few solutions that could help us in countering attacks (Matthews, 2019). For example, networks could contain intrusion detection systems (IDSs), which warn defenders about potential cyberattacks (Bace and Mell, 2001; Aggarwal et al., 2018; Aggarwal and Dutt, 2020). Although robust, IDSs may suffer from false alarms (indicating a cyber-threat when one is not present) and misses (missing to show a cyber-threat when it is present) (Mell et al., 2003). These false alarms and misses could lead to loss of revenue and significant damages to cyberinfrastructure, respectively (Shang, 2018a). Prior research has also proposed that hybrid censoring and filtering strategies may enable bounded non-rational network agents to reach consensus behavior (Shang, 2018b, 2019). Overall, such consensus could be useful in detecting cyberattacks before they become damaging (Shang, 2018b).

Beyond IDSs and filtering strategies, another solution that has been shown to be effective against cyberattacks is deception (Cohen, 2006; Aggarwal et al., 2016a; Almeshekah and Spafford, 2016; Dutt et al., 2016). In fact, deception via honeypots (systems that pretend to be real) has been a prominent technique for the detection, prevention, and response to cyberattacks (Garg and Grosu, 2007; Rowe and Custy, 2007; Heckman et al., 2013; Aggarwal et al., 2016a,b; Almeshekah and Spafford, 2016). In the real world, such honeypots may be created via port hardening or by putting fake content in computer systems (Shimeall and Spring, 2013). Deception via honeypots has also been used in cutting-edge technologies like the Internet of things (IoT) to defend against modern cyberattacks (La et al., 2016).

Some researchers have proposed games to study the role of deception in cybersecurity mathematically (Garg and Grosu, 2007; Kiekintveld et al., 2015; Aggarwal et al., 2017). However, more recently, researchers have investigated human decisions in the presence of deception in abstract Stackelberg security games (Cranford et al., 2018) as well as applied games like HackIT (Aggarwal et al., 2019, 2020). Here, researchers have relied upon behavioral game theory (Camerer, 2003) and cognitive theories like instance-based learning theory (IBLT) (Gonzalez et al., 2003; Gonzalez and Dutt, 2011, 2012; Dutt and Gonzalez, 2012; Dutt et al., 2013) to understand human decisions in different cyberattack scenarios (Aggarwal et al., 2020).

Human decisions in different cyberattack scenarios may be influenced by a host of different factors, including variety and complexity of cyberattacks, network topology, and the number and diversity of zero-day vulnerabilities (Garcia-Teodoro et al., 2009; Wang et al., 2010; Lenin et al., 2014). One factor that has been less investigated and that is likely to influence human decisions in cyberattack scenarios is the network size (i.e., the number of computer systems in the network; Bagchi and Tang, 2004; Wang et al., 2010). For example, Bagchi and Tang (2004) demonstrated via computational modeling that network size was an influencing factor in different kinds of cyberattacks. Similarly, as per Bagchi and Tang (2004) and Wang et al. (2010), as the size of the network increases, one expects growth in the proportion of cyberattacks. Although prior research has investigated the influence of network size on cyberattacks via computational modeling, very little is known on how the size of the network influences human adversarial decisions in games involving deception.

Thus, the primary objective of this research is to understand the influence of network size on human adversarial decisions in games involving deception. Specifically, we develop a novel cybersecurity game involving deception via honeypots, and we vary the number of computer systems in a simulated network in the game across different experimental conditions. In the deception game (DG), adversaries can first probe some of the computer systems and then decide what systems to attack for real. In a network of different sizes, the proportion of honeypots remains constant. The outcomes of this research may help cybersecurity professionals in understanding the robustness of the honeypot network architectures of varying sizes against modern cyberattacks.

In what follows, first, we detail a DG and how the network size was varied in this game. Next, we state our expectations on the influence of network size on decisions in DG using IBLT. Furthermore, we test these expectations in an experiment involving human participants. Finally, we evaluate the results from the experiment and highlight their implications for using deception in the real world.

THE DG

The DG is a sequential, single-player game, i.e., a game between an adversary and a network (Garg and Grosu, 2007; Aggarwal et al., 2016a,b). The game is formally denoted as DG (n, k, γ), where n is the total number of web servers, k is the number of honeypots, and γ is the number of probes after which the adversary makes his final decision to attack the network or not and γ should be less than or equal to k (Garg and Grosu, 2007). There are two kinds of web servers in the game, regular and honeypot. Regular web servers are the real web servers, which contain valuable information, whereas honeypots are fake servers, which pretend to be regular with the aim of trapping adversaries to extract meaningful information. The objective of the adversary is to attack the regular web server and gain maximum points.

The game is played for multiple rounds. In each round of this game, we have two stages, the probe stage and attack stage. In the probe stage, an adversary could probe web servers multiple times. Probing means clicking on the button which denotes a web server in the game's interface. For each probe, the adversary gets a response from the system about the system being a regular (real) web server or a honeypot (fake) web server. This feedback may or may not be accurate depending on the absence or presence of deception, respectively. Thus, this scenario may not allow the adversary to learn across a number of rounds of play. Furthermore, the game dynamics may likely mimic the real world, where adversaries may only have limited information about the nature of the infrastructure they are trying to compromise. Overall, the purpose of deception is to fool the adversary by making her believe in false information about the state of the servers. If deception is present in a round, then the network response is opposite the actual state of web servers. Thus, if the adversary probes a regular web server, then the network's response is "honeypot," and if the adversary probes a honeypot, then the network's response is "regular." If deception is not present in a round, then the network's response will be the same as the actual state of web servers. Thus, if the adversary probes a regular web server, the network's response is "regular," and if the adversary probes a honeypot web server, the network's response is "honeypot." In the probe stage, the adversary has an additional option not to probe any web server. Deception and unreliability in feedback of the probe stage might increase no-attack actions, as the unreliable feedback of the probe stage will likely make the adversary avoid risk for regular/honeypot attack actions.

Once the adversary has made γ number of independent probes (or decides not to probe any web server), the game enters the attack stage. In the attack stage, the adversary decides to attack one of the web servers once. Attacking means clicking on the button which denotes a web server. The adversary may also decide not to attack any web server in the attack stage. Based upon the decisions made during the probe and attack stages, the adversary may win or lose points. **Table 1** shows the payoff matrix for the adversary based upon the decisions in the probe and attack stages in the DG.

As shown in **Table 1**, in each round, if the adversary probes/attacks a regular web server, then the adversary is awarded positive points. If the adversary probes/attacks a honeypot web server, then the adversary is awarded negative points. If the adversary does not probe/attack any web server in any of the rounds, he neither loses nor gains any points. Thus, if the adversary probes a regular web server, he gains +5 points, whereas on probing a honeypot web server, he loses -5 points. If the adversary attacks a regular web server, he gains +10 points, whereas he loses -10 points on probing a honeypot web server. After completion of the attack stage, the total score of a round is calculated; and at the end of the multiple rounds, the cumulative score is calculated. The values of the payoff in **Table 1** were motivated from prior literature (Aggarwal et al., 2016a,b).

INFLUENCE OF NETWORK SIZE ON ADVERSARY'S DECISION

In our experiment, there were three different versions of the DG to simulate networks of different sizes. Motivated from networks in the real world, the versions of the game included DG (2, 1, 1) (small), DG (6, 3, 3) (medium), and DG (12, 6, 6) (large). We kept the proportion of honeypots to the total number of web servers constant (at 50%) across the three versions of the game. Also, the number and sequence of deception and non-deception rounds were kept the same for all three versions of the DG.

TABLE 1 | Adversary's payoff during the probe stage and attack stage in the DG.

Stage	Adversary's Action	Adversary's Payoff
Probe	Regular web server	+5 points
	Honeypot web server	−5 points
	Do not probe	0 points
Attack	Regular Web server	+10 points
	Honeypot web server	−10 points
	Do not attack	0 points

Though the proportion of honeypots is the same across all three network sizes, we expect adversaries to probe and attack regular and honeypot web servers much less in the small-sized network compared to medium- or large-sized networks. One could explain this expectation based upon cognitive theories like IBLT (Gonzalez et al., 2003; Gonzalez and Dutt, 2011, 2012; Dutt and Gonzalez, 2012; Dutt et al., 2013). As per IBLT, human decisions may be driven by the exploration of available options during information search (probing) and their exploitation during real decisions (attack). Decision making during different probe and attack stages will be likely determined in a bounded-rational manner by reliance on recency and frequency of decision and their outcomes (i.e., human decisions will be driven by forgetting of distant instances and recall of only recent instances). When the network size is small, the decisions during probe and attack stages in DG involve a choice between two web servers, where one of them is a honeypot. Given the smaller number of web servers, it may be easier for bounded-rational decision makers to recall the mapping of web servers being regular or honeypot from memory. That is because fewer instances will be created in memory corresponding to the different web servers, and their activations will be high in memory due to smaller delays in their exploration during probing. However, in the medium- and large-sized networks, due to the presence of multiple web servers, bounded-rational decision makers may not be able to easily recall the mapping of web servers as regular or honeypot from memory. That is because multiple instances, one per web server, will be created in memory, and the activation of these instances will likely not be high in memory due to the long delays in their exploration during probing. Overall, the difficulty in the recall of distant instances in medium- and large-sized networks may cause more exploration of web servers during the probe stage and the attack stage in these configurations compared to that in the small-sized network. Thus, based upon IBLT, one expects that the proportion of probe and attack actions on regular and honeypot web servers will be more in medium- and large-sized networks compared to the proportion of probe and attack actions in the small-sized network. Furthermore, as instances corresponding to no-probe and no-attack actions will be more activated in memory in the small-sized network compared to medium- and large-sized networks, we expect a larger proportion of these no-probe and no-attack actions in the small-sized network compared to medium- and large-sized networks. That is because no-probe and no-attack instances in memory will be easier to recall in a small-sized network compared to medium-sized or large-sized networks. Next, we test these expectations based upon IBLT in an experiment involving human decision makers making decisions in DG.

EXPERIMENT

In this section, we detail the experiment we carried out with human participants performing as adversaries across all rounds in the DG. The game was used to calculate the effectiveness of honeypots in different-sized networks.

Methods

Experiment Design

Participants performing as an adversary ("hacker") were randomly assigned to one of three between-subjects network size conditions (*N* = 20 participants per condition): DG (2, 1, 1) (small), DG (6, 3, 3) (medium), and DG (12, 6, 6) (large). Each condition in DG was 29 rounds long, where there were 14 deception rounds and 15 non-deception rounds (participants did not know what rounds were deception rounds and what rounds were non-deception rounds). The sequence of the deception and non-deception rounds was randomized once and then kept the same across all three conditions (see the **Supplementary Material** for the sequence of deception and non-deception rounds). In a round, the assignment of honeypots and regular web servers to buttons was done randomly. In the small network, the DG involved two web servers, where one of them was randomly assigned as a honeypot, and the adversary could probe one of the web servers in the probe stage (the adversary may also decide not to probe any of the web servers). In the medium network, the DG involved six web servers, where three web servers were randomly selected to be honeypots, and the adversary could probe web servers three times in the probe stage (the adversary may also decide not to probe any of the web servers). In the large network, the DG involved 12 web servers, where six web servers were randomly selected to be honeypots, and the adversary could probe the web servers six times in the probe stage (the adversary may also decide not to probe any of the web servers). Across all network sizes, after completion of the probe stage, the adversary entered the attack stage. If the adversary decided not to probe a web server anytime during the probe stage, then the probe stage ended, and the adversary entered the attack stage. In the attack stage, the adversary either decided to attack one of the web servers or decided not to attack any of them. In each condition, dependent measures included regular probe/attack proportions, honeypot probe/attack proportions, and no-web server probe/attack proportions. For computing these proportions, each regular probe/attack action by a participant in a round was coded as rp/ra, each honeypot probe/attack action by a participant in a round was coded as hp/ha, and no-web server probe/attack action was coded as np/na. Later, we computed the proportions as rp/Tp, ra/Ta, hp/Tp, ha/Ta, np/Tp, and na/Ta, where Tp and Ta were the total number of decisions during probe and attack stages, respectively, in a condition. Later, these proportions were averaged across all participants in a condition.

Stimuli

Figure 1 shows the interface shown to participants in the probe stage of the DG with six web servers. As shown in the figure, participants were informed about the task with short instructions regarding the different types of web servers. Once the participant probed one of the web servers by clicking the corresponding button, she received the response from the web server (see **Figure 2**). Once the participant had probed for a fixed number of times, she proceeded to the attack stage (see **Figure 3**). After attacking one of the web servers in the network, the participant's score was calculated for the round based on his actions in the probe and attack stages (see **Figure 4**).

Participants

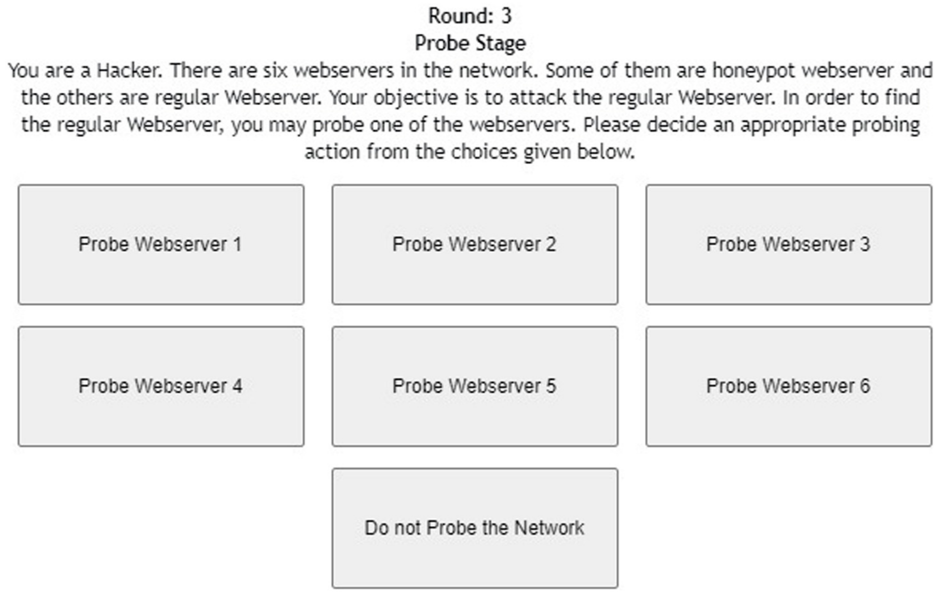
This study was conducted after approval of the Ethics Committee at the Indian Institute of Technology Mandi (IITM/DST-ICPS/VD/251) with written consent from all participants. Participation was voluntary, and all participants gave written consent before starting their study. Participants were anonymously recruited for the cybersecurity study through the Amazon Mechanical Turk, a crowdsourcing website (Mason and Suri, 2012). Eighty-six percent of participants were male, and the rest were females. The age of participants ranged between 19 and 48 years (median = 31 years, mean = 32, and standard deviation = 6 years). Around 92% of participants possessed a college degree, while the remaining 8% were currently pursuing a college degree. Also, 60% of the participants had science, technology, engineering, and mathematics as a major. Participants were paid a participation fee INR 50 (USD 0.7) after they completed their study. The top three scorers of the game were chosen for the lucky draw contest, and one of these participants was randomly selected for a gift voucher of INR 500 (USD 7.14). The score was computed based upon points earned in the game during the probe and attack stages across 29 rounds.

Procedure

Participants performing as adversaries were given instructions about their goal in DG. Participants were told that there might be deception present in DG with both regular and honeypot servers; however, participants were not told which exact web servers were regular and which were honeypots. Participants were asked to maximize their score across several rounds involving the probe and attack stages in DG, but the endpoint in the study was not disclosed to participants. Each round has two stages: the probe stage and the attack stage. An adversary could probe multiple web servers in the DG for medium and large networks, whereas she could probe only one web server in a small network. In all network size conditions, adversaries could attack only one of the web servers in the attack stage. Once the study was completed, participants were thanked and paid for their participation. A copy of the instructions from one of the conditions is provided as **Supplementary Material**.

Data Analyses

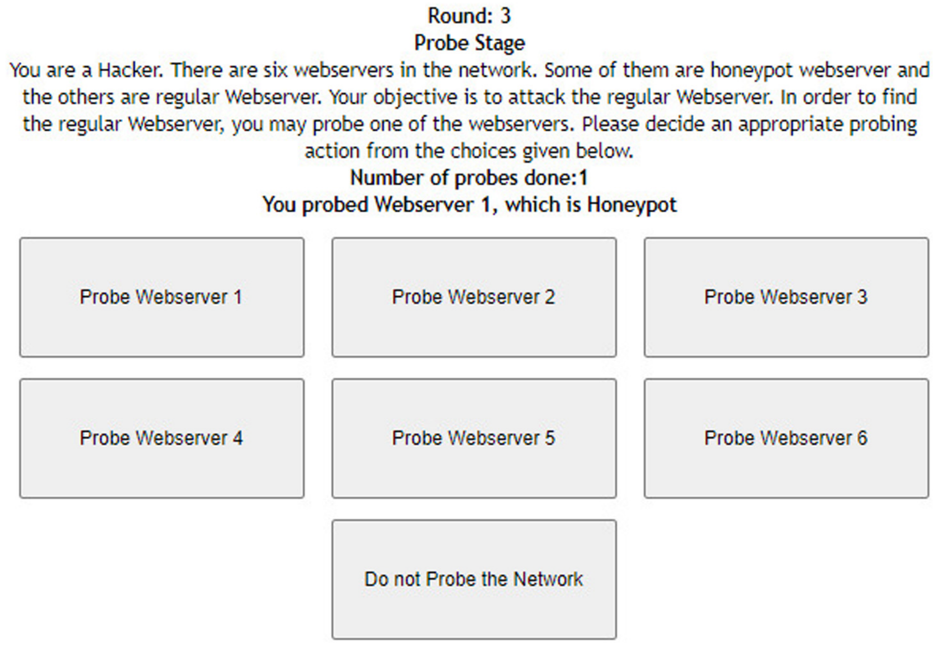
We used analysis of variance (ANOVA), a statistical technique, to test differences between two or more means across different network size conditions (Field, 2013). Also, as sample sizes were equal across different conditions, we used the Tukey *post hoc* test (Field, 2013). The alpha level or the *p*-value (the probability of rejecting the null hypothesis when it is true) was set at 0.05, and power (the probability of rejecting the null hypothesis when it is false) was set at 0.80. We performed one-way ANOVAs to investigate the influence of network size



Round: 3
Probe Stage

You are a Hacker. There are six webservers in the network. Some of them are honeypot webserver and the others are regular Webserver. Your objective is to attack the regular Webserver. In order to find the regular Webserver, you may probe one of the webservers. Please decide an appropriate probing action from the choices given below.

FIGURE 1 | Probe stage of the deception game with six web servers.



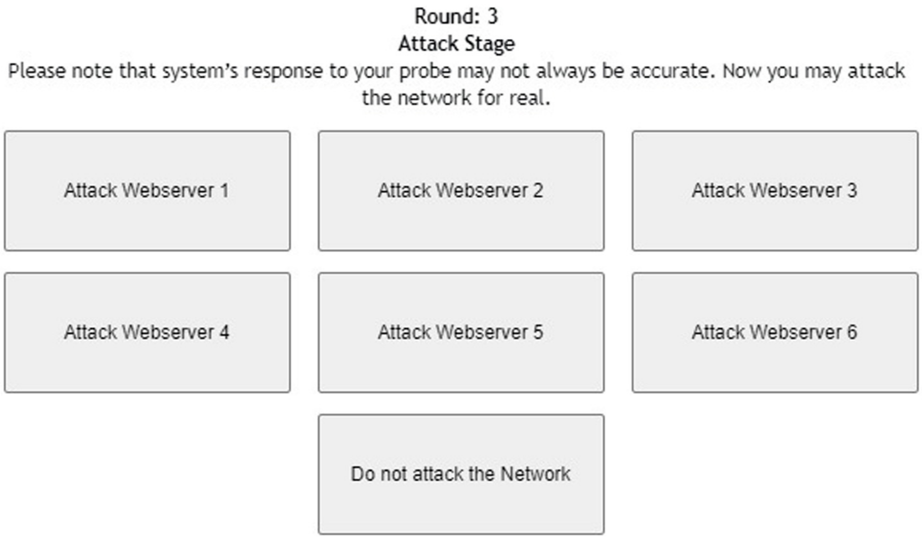
Round: 3
Probe Stage

You are a Hacker. There are six webservers in the network. Some of them are honeypot webserver and the others are regular Webserver. Your objective is to attack the regular Webserver. In order to find the regular Webserver, you may probe one of the webservers. Please decide an appropriate probing action from the choices given below.

Number of probes done:1
You probed Webserver 1, which is Honeypot

FIGURE 2 | Probe stage of the deception game with six web servers after the participant probes for the first time.

on regular attack, honeypot attack, and no-attack decisions during the probe and attack stages. Also, we performed two-way mixed-factorial ANOVAs with network size as a between-subjects factor and sequential probe-attack trials as a within-subjects factor. Based upon the Q–Q plots (between expected quantiles and normal quantiles), different dependent variables (regular probe/attack decisions, honeypot probe/attack decisions, and no-probe/attack decisions) were found to be normally distributed. Similarly, Levene's test showed that the variances were homogeneous for different decisions during both the probe and attack stages: honeypot web server probe [$F(2,57) = 0.641$, $p = 0.53$], regular web server probe [$F(2,57) = 1.22$, $p = 0.30$], no web server probe [$F(2,57) = 0.382$, $p = 0.68$], regular web server attack [$F(2,57) = 2.11$, $p = 0.13$], honeypot web server attack [$F(2,57) = 1.19$, $p = 0.31$], and no web server attack [$F(2,57) = 3.70$, $p = 0.07$].

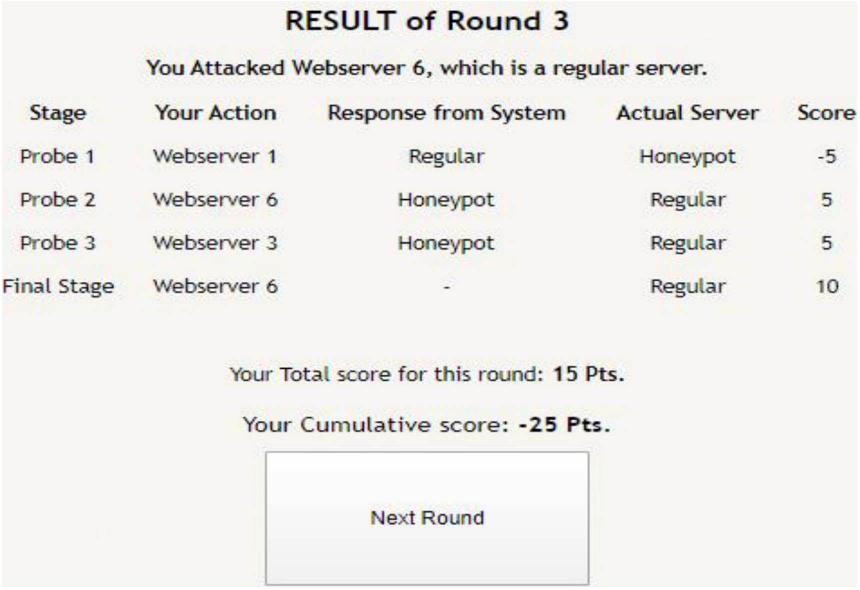


Round: 3

Attack Stage

Please note that system's response to your probe may not always be accurate. Now you may attack the network for real.

FIGURE 3 | Attack stage of the deception game with six web servers.



RESULT of Round 3

You Attacked Webserver 6, which is a regular server.

Stage	Your Action	Response from System	Actual Server	Score
Probe 1	Webserver 1	Regular	Honeypot	-5
Probe 2	Webserver 6	Honeypot	Regular	5
Probe 3	Webserver 3	Honeypot	Regular	5
Final Stage	Webserver 6	-	Regular	10

Your Total score for this round: **15 Pts.**

Your Cumulative score: **-25 Pts.**

FIGURE 4 | Result of a completed round, where a participant gets to know his score based upon his actions in the probe and attack stages.

RESULTS

Descriptive Statistics

In our experiment, we had three different dependent variables in the probe and attack stages in the DG. In the probe stage, we had a regular web server probe, honeypot web server probe, and no web server probe. Similarly, in the attack stage, we had a honeypot web server attack, regular web server attack, and no web server attack. **Table 2** describes the descriptive statistics for different dependent variables in the experiment across all conditions.

Influence of Network Size on Decisions During the Probe Stage

We performed one-way ANOVA to investigate the influence of network size on decisions during the probe stage. The network size significantly influenced the proportion of honeypot web server probes [$F(2,59) = 35.86$, $p < 0.001$, $\eta^2 = 0.56$], regular web server probes [$F(2,59) = 18.31$, $p < 0.001$, $\eta^2 = 0.39$], and no web server probes [$F(2,59) = 34.39$, $p < 0.001$, $\eta^2 = 0.55$], where p-value tests the statistical significance in the hypothesis test and η^2 denotes the measure of the effect size. **Figure 5** shows

TABLE 2 | Descriptive statistics for different dependent variables in the experiment.

Stage	Dependent Variable	Mean	Std. Deviation	Minimum	Maximum
Probe	Honeypot web server probe	0.38	0.18	0.03	0.56
	Regular web server probe	0.39	0.13	0.06	0.57
	No web server probe	0.23	0.27	0.00	0.91
Attack	Honeypot web server attack	0.40	0.18	0.07	0.66
	Regular web server attack	0.40	0.12	0.17	0.69
	No web server attack	0.20	0.25	0.00	0.69

the proportion of honeypot, regular, and no web server probes across different network sizes.

As shown in **Figure 5**, the proportion of honeypot web server probes was 0.22 in the small network; however, the proportions of honeypot web server probes were 0.45 and 0.47 in the medium and large networks, respectively. The Tukey *post hoc* tests revealed that the proportion of honeypot web server probes in the small network was significantly smaller compared to the proportions of honeypot web server probes in the medium network ($p < 0.001$) and large network ($p < 0.001$). However, as per the Tukey *post hoc* tests, there were no significant differences between the proportions of honeypot web server probes in the medium and large networks ($p = 0.83$). These results are as per our expectations.

As shown in **Figure 5**, the proportion of regular web server probes was 0.27 in the small network; however, the proportions of regular web server probes were 0.45 and 0.45 in the medium and large networks, respectively. The Tukey *post hoc* tests revealed that the proportion of regular web server probes in the small network was significantly smaller compared to the proportions of regular web server probes in the medium network ($p < 0.001$) and large network ($p < 0.001$). However, as per the Tukey *post hoc* tests, there was no significant difference between the proportions of regular web server probes in the medium and large networks ($p = 0.99$). These results are as per our expectations.

As shown in **Figure 5**, the proportion of no web server probes was 0.51 in the small network; however, the proportions of no web server probes were 0.10 and 0.08 in the medium and large networks, respectively. The Tukey *post hoc* tests revealed that the proportion of no web server probes in the small network was significantly smaller compared to the proportions of no web server probes in the medium network ($p < 0.001$) and large network ($p < 0.001$). However, as per the Tukey *post hoc* tests, there was no significant difference between the proportions of no web server probes in the medium and large networks ($p = 0.92$). These results are as per our expectations.

Influence of Network Size on Decisions During Attack Stage

We performed one-way ANOVAs to investigate the influence of network size on decisions during the attack stage. The network size significantly influenced the proportion of honeypot web server attacks [$F(2,59) = 51.77$, $p < 0.001$, $\eta^2 = 0.65$], regular

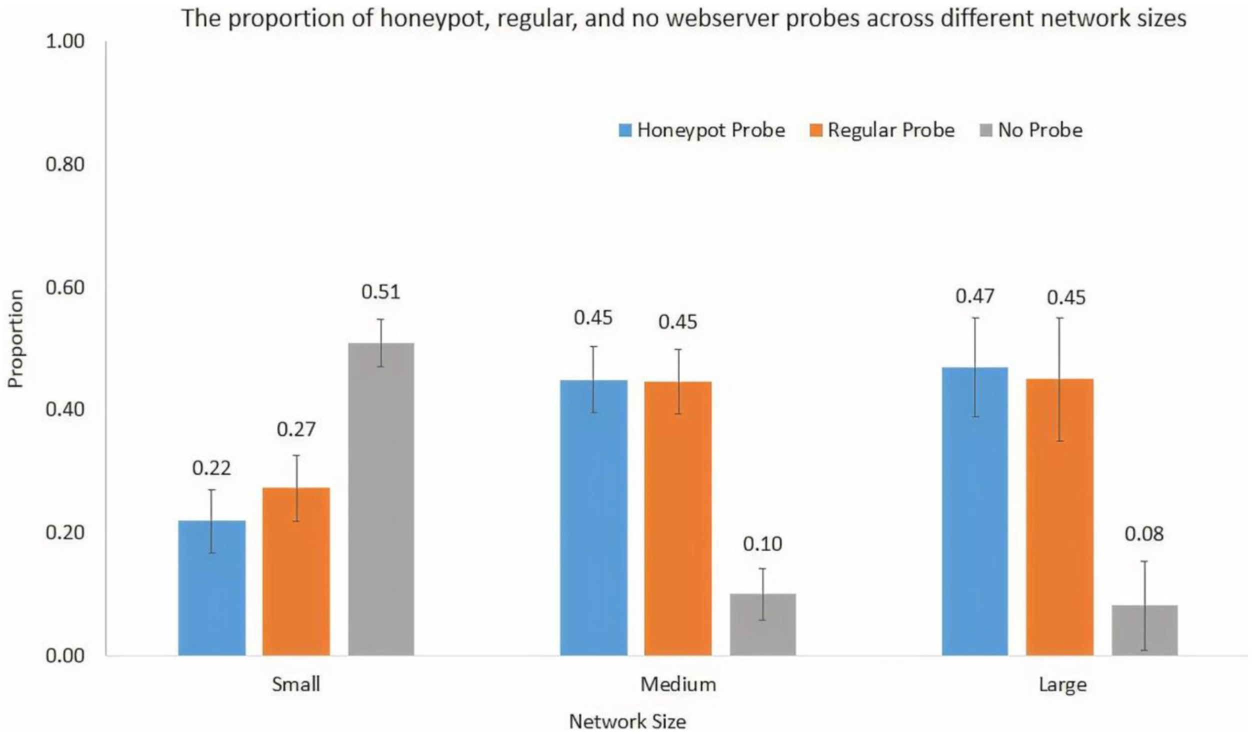


FIGURE 5 | The proportion of honeypot probe, regular probe, and no-probe decisions across different network sizes.

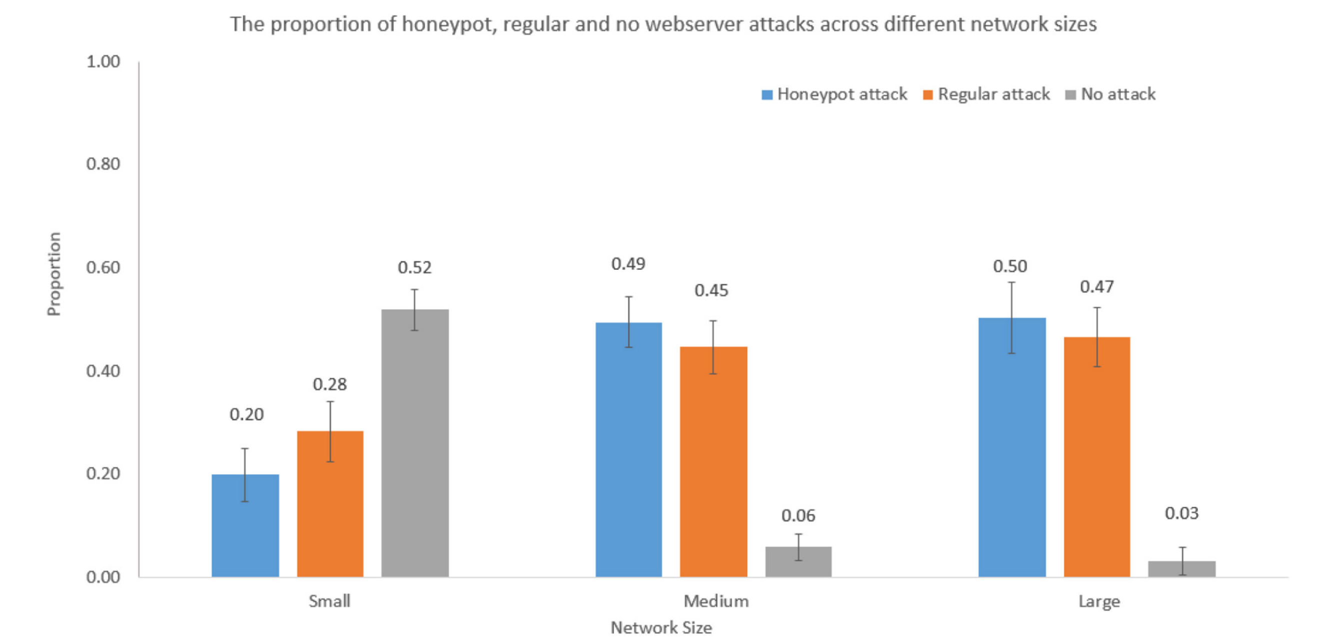


FIGURE 6 | The proportion of honeypot attack, regular attack, and no-attack decisions across different network sizes.

web server attacks [$F(2,59) = 23.32$, $p < 0.001$, $\eta^2 = 0.45$], and no web server attacks [$F(2,59) = 111.68$, $p < 0.001$, $\eta^2 = 0.78$]. **Figure 6** shows the proportion of honeypot, regular, and no web server attacks across different network sizes.

As shown in **Figure 6**, the proportion of honeypot web server attacks was 0.20 in the small network; however, the proportions of honeypot web server attacks were 0.49 and 0.50 in the medium and large networks, respectively. The Tukey *post hoc* tests revealed that the proportion of honeypot web server attacks in the small network was significantly smaller compared to the proportions of honeypot web server attacks in the medium network ($p < 0.001$) and large network ($p < 0.001$). However, as per the Tukey *post hoc* tests, there were no significant differences between the proportion of honeypot web server attacks in the medium and large networks ($p = 0.97$). These results are as per our expectations.

As shown in **Figure 6**, the proportion of regular web server attacks was 0.28 in the small network; however, the proportions of regular web server attacks were 0.45 and 0.47 in the medium and large networks, respectively. The Tukey *post hoc* tests revealed that the proportion of regular web server attacks in the small network was significantly smaller compared to the proportions of regular web server attacks in the medium network ($p < 0.001$) and large network ($p < 0.001$). However, as per the Tukey *post hoc* tests, there was no significant difference between the proportion of regular web server attacks in the medium and large networks ($p = 0.80$). These results are as per our expectations.

As shown in **Figure 6**, the proportion of no web server attacks was 0.52 in the small network; however, the proportions of no web server attacks were 0.06 and 0.03 in the medium and large networks, respectively. The Tukey *post hoc* tests revealed that the proportion of no web server attacks in the small network was significantly smaller compared to the proportions of no web server attacks in the medium network ($p < 0.001$) and large network ($p < 0.001$). However, as per the Tukey *post hoc* tests, there was no significant difference between the proportion of no web server attacks in the medium and large networks ($p = 0.73$). These results are as per our expectations.

Influence of Network Size and Sequential Probe/Attack Trials on Decisions

We performed mixed-factorial ANOVAs with network size as a between-subjects factor and sequential probe/attack trials as a within-subjects factor. The network size significantly interacted with sequential probe/attack trials for the following decisions: honeypot server probed and no server attacked [$F(2,57) = 91.92$, $p < 0.001$, $\eta^2 = 0.76$]; regular server probed and honeypot server attacked [$F(2,57) = 6.40$, $p < 0.001$, $\eta^2 = 0.18$]; regular server probed and no server attacked [$F(2,57) = 81.23$, $p < 0.001$, $\eta^2 = 0.74$]; no server probed and regular server attacked [$F(2,57) = 49.29$, $p < 0.001$, $\eta^2 = 0.63$]; no server probed and honeypot server attacked [$F(2,57) = 54.15$, $p < 0.001$, $\eta^2 = 0.66$].

Figure 7 shows the interaction between network size and honeypot server probed and no server attacked decisions. For a small network, the proportion of honeypot server probed was 0.22, and the proportion of no server attacked decisions was 0.52. However, for medium and large networks, the proportions of honeypot server probed were 0.45 and 0.47, and the proportions of no server attacked decisions were 0.06 and 0.03, respectively.

Figure 8 shows the interaction between network size and regular server probed and honeypot server attacked decisions. For a small network, the proportion of regular server probed was 0.27, and the proportion of honeypot server attacked

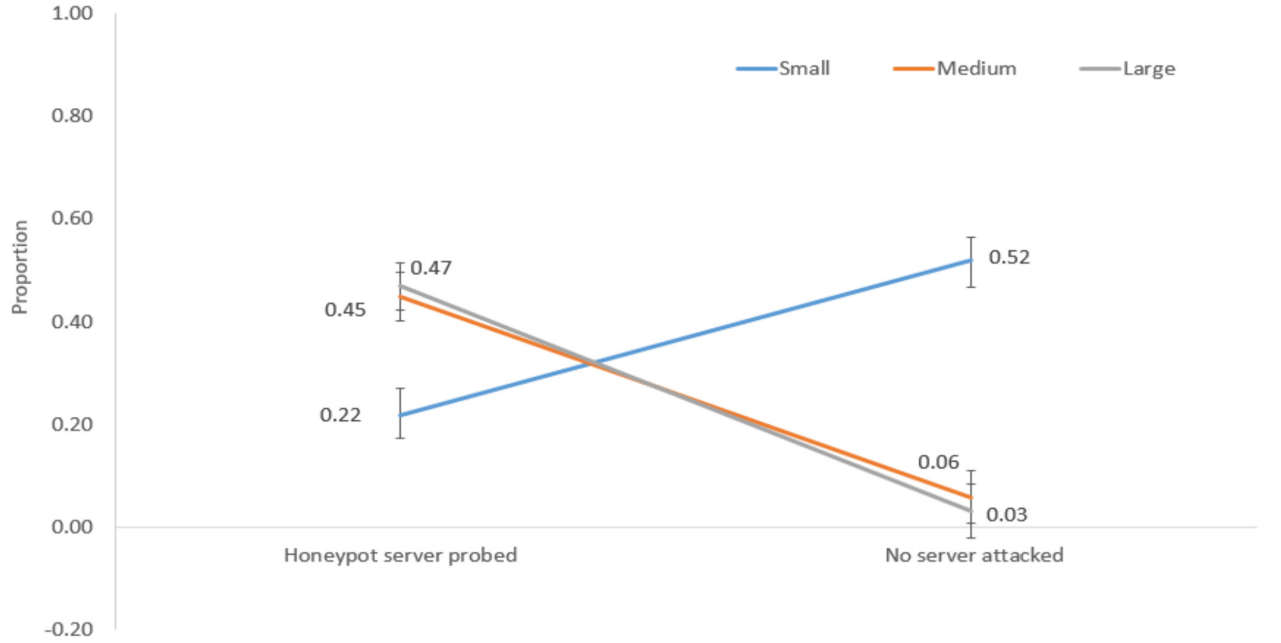


FIGURE 7 | The proportions of honeypot server probed and no server attacked decisions in different network sizes.

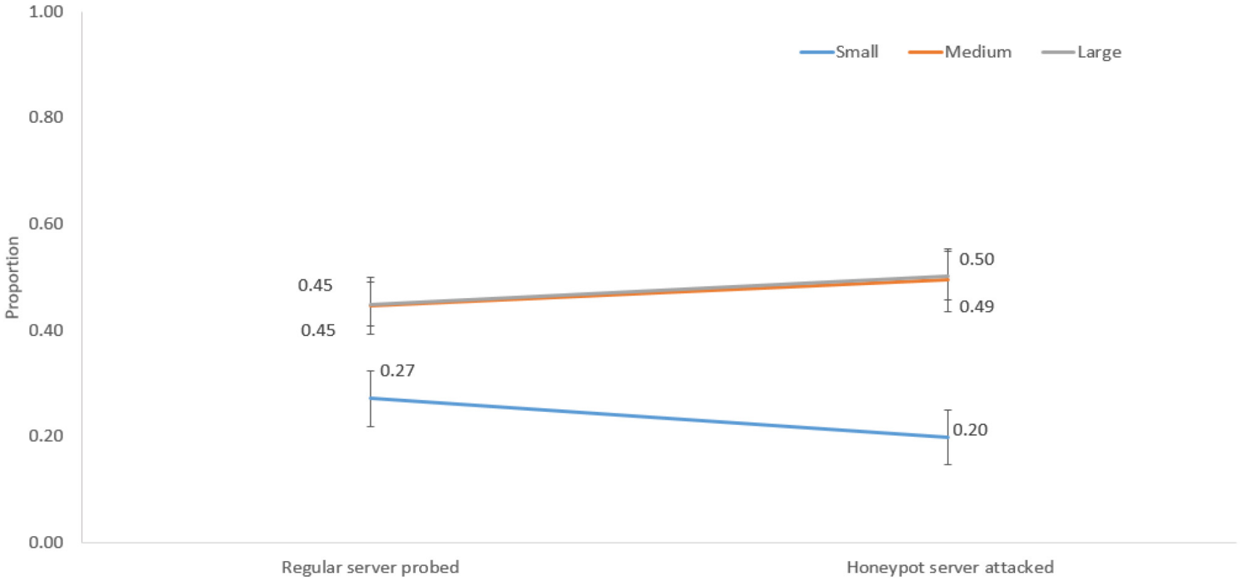


FIGURE 8 | The proportion of regular server probed and honeypot server attacked decisions in different network sizes.

decisions was 0.20. However, for medium and large networks, the proportions of regular server probed were 0.45 and 0.45, and the proportions of honeypot server attacked decisions were 0.49 and 0.50, respectively.

Figure 9 shows the interaction between network size and regular server probed and no server attacked decisions. For a small network, the proportion of regular server probed was 0.27, and the proportion of no server attacked decisions was 0.52. However, for medium and large networks, the proportions of regular server probed were 0.45 and 0.45, and the proportions of no server attacked decisions were 0.06 and 0.03, respectively.

Figure 10 shows the interaction between network size and no server probed and regular server attacked decisions. For a small network, the proportion of no server probed was 0.51, and the proportion of regular server attacked decisions was 0.28. However, for medium and large networks, the proportions of no server probed were 0.10 and 0.08, and the

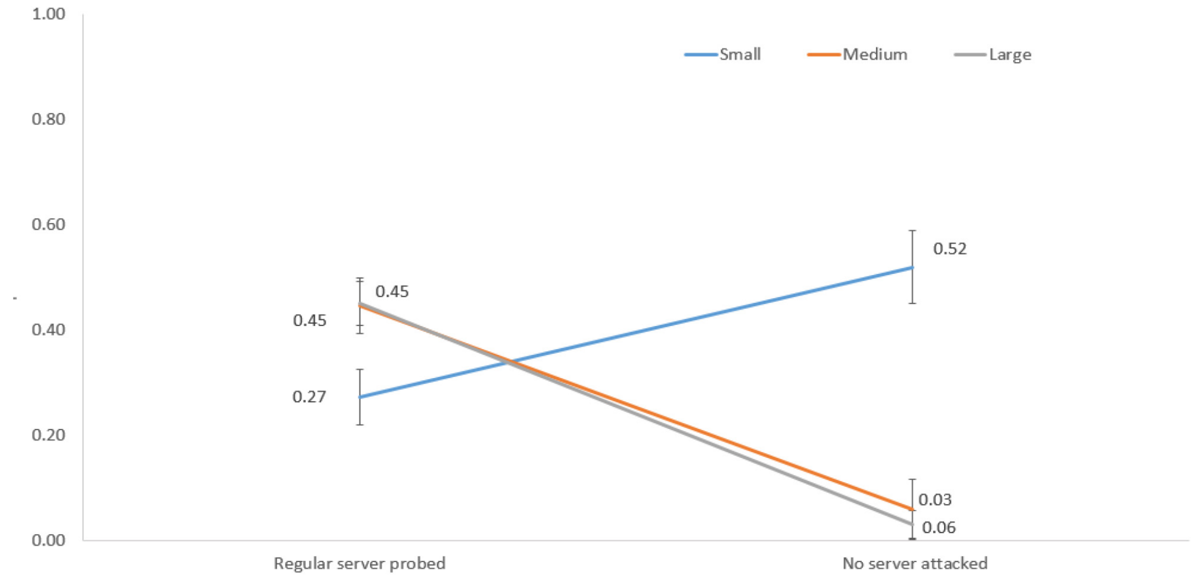


FIGURE 9 | The proportions of regular server probed and no server attacked decisions in different network sizes.

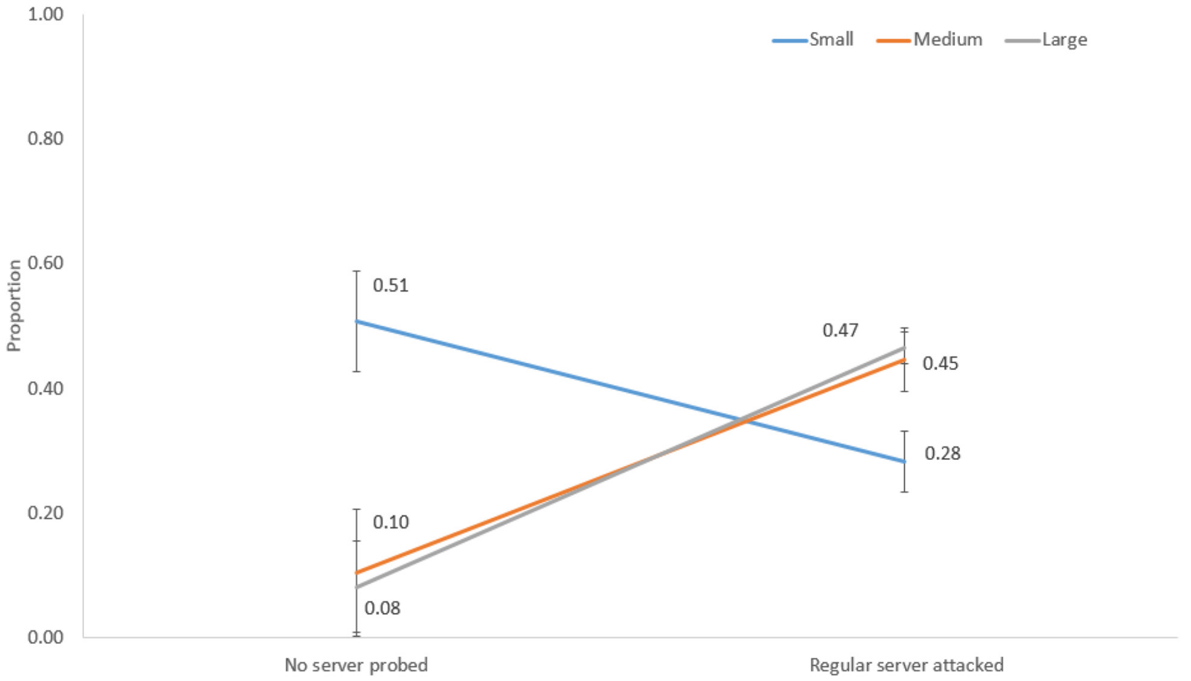


FIGURE 10 | The proportions of no server probed and regular server attacked decisions in different network sizes.

proportions of regular server attacked decisions were 0.45 and 0.47, respectively.

Figure 11 shows the interaction between network size and no server probed and the honeypot server attacked decisions. For a small network, the proportion of no server probed was 0.51, and the proportion of honeypot server attacked decisions was 0.20. However, for medium and large network sizes, the proportions of no server probed were 0.10 and 0.08, and the proportions of honeypot server attacked decisions were 0.49 and 0.50, respectively.

DISCUSSION AND CONCLUSION

Deception via honeypots can act as an essential tool to defend cyberattacks (Cohen, 2006; Rowe and Custy, 2007). Although

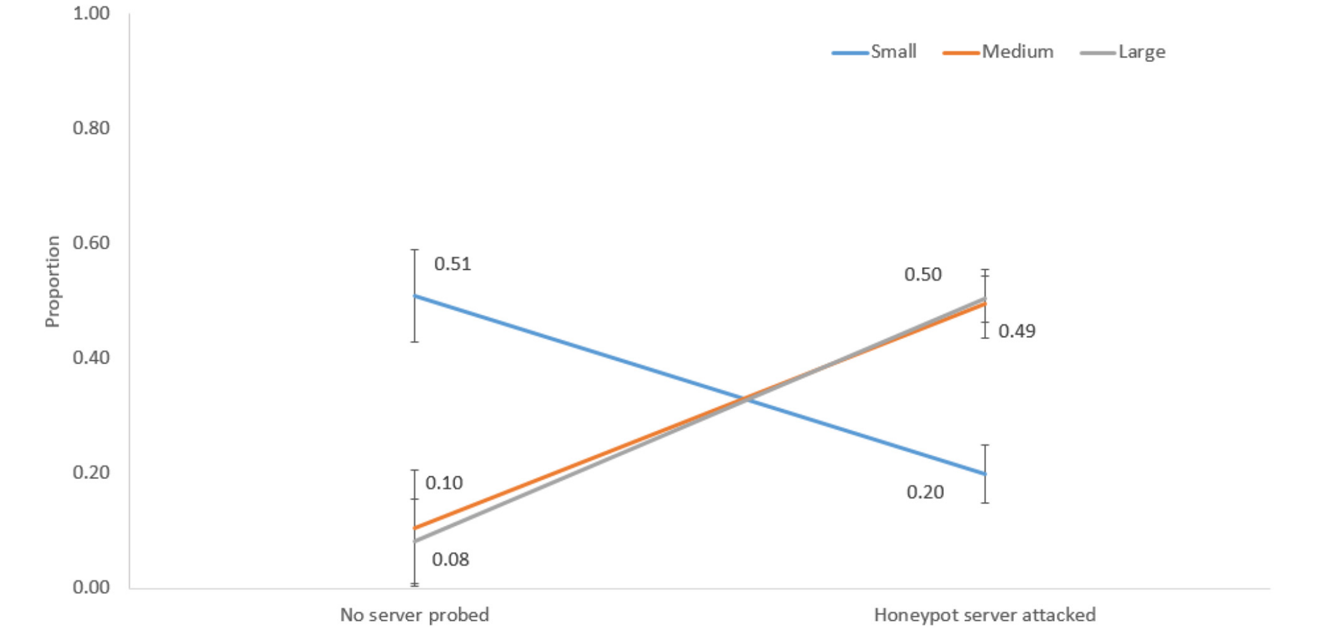


FIGURE 11 | The proportions of no server probed and honeypot server attacked decisions in different network sizes.

prior research has developed and used games to understand the role of deception in cybersecurity, researchers had yet to investigate how the network's size (i.e., the number of computers on the network) influences the adversary's probe and attack decisions in the presence of deception via honeypots. To address this gap in the literature, in this paper, we investigated the influence of network size on adversary's decisions in a DG involving honeypot web servers. Results revealed that the proportions of honeypot probe and attack actions and the proportions of regular probe and attack actions were more in medium- and large-sized networks compared to small-sized networks. Also, there was an influence of probing actions on attack actions across all three network sizes. These results can be explained based upon the IBLT, a theory of decisions from experience (Gonzalez et al., 2003; Gonzalez and Dutt, 2011, 2012; Dutt et al., 2013).

First, results revealed that the proportions of honeypot and regular probes and attacks were more in medium- and large-sized networks compared to small-sized networks. When the network size is small, the decisions during probe and attack stages in DG involve a choice between two web servers, where one of them is a honeypot. Given the smaller number of web servers, as per IBLT, it may be easier for bounded-rational participants to recall the mapping of web servers being regular or honeypot from memory. That is because about two instances are created in memory when there are two web servers, and the activation of these instances is likely to be much higher in memory due to smaller delays in their exploration during probing. However, in medium- and large-sized networks, due to the presence of multiple web servers, bounded-rational participants may not be able to easily recall the mapping of web servers as regular or honeypot from memory. That is because multiple instances, one per web server, would be created in memory, and the activation of these instances will likely decay in memory due to the long delays in their exploration during probing. Overall, as per IBLT, the difficulty in the recall of distant instances in medium- and large-sized networks may cause more exploration of web servers during the probe stage and the attack stage in these configurations compared to that in the small-sized network.

Second, the proportions of no-probe and no-attack actions were more in small-sized networks compared to medium- and large-sized networks. As per IBLT, a likely reason for these results is the differential activation of instances in memory for the no-probe and no-attack actions across the different-sized networks. As there would be fewer instances created in memory in small-sized networks compared to medium- and large-sized networks, these smaller numbers of memory instances corresponding to no-probe and no-attack actions are likely to be more activated in the small-sized network compared to medium- and large-sized networks. Overall, due to their higher activations, the no-probe and no-attack instances in memory will be easier to recall in a small-sized network compared to medium- or large-sized networks.

Third, we investigated the influence of network size and sequential probe/attack trials in DG. First, probing a honeypot caused an increase (decrease) in no server attacked actions in small (medium or large) networks. Second, probing a regular server caused a decrease (increase) in honeypot server attacked actions and an increase (decrease) in no server attack actions in small (medium or large) networks. Third, not probing a server caused a decrease (increase) in regular and honeypot server attacks in small (medium or large) networks. All these results can be explained based upon the differences in the activation and number of instances in memory in small-sized networks compared to large-sized networks. For example, as there were fewer and more activated instances likely created in memory of participants playing in small-sized networks compared to medium- and large-sized networks, the decisions of participants in small-sized networks were more logical and deterministic compared to those playing in medium- and large-sized networks. Due to these differences, perhaps, it was reasonable for participants playing in a small-sized network to show the above-stated results. At the same time, due to larger and weakly activated instances in memory of those playing medium- and large-sized networks, their decisions seemed to be less logical and more exploratory.

In this research, we performed a laboratory experiment using a canonical game, and our conclusions should be seen with this assumption in mind. However, our results have some important implications for the real world. First, our results reveal that making networks larger has an effect of increasing the proportion of regular probes and regular attacks. Thus, it may be advisable to break larger networks into smaller subnetworks, where these subnetworks may only possess a subset of computers (Achleitner et al., 2017). Furthermore, if these smaller subnetworks possess a number of honeypots, then these honeypots will likely cause adversaries to encounter them and not to attack the network. Also, a decrease in probes in these subnetworks may likely cause a decrease in the number of regular attacks.

One limitation of our research is that our results are derived from a lab-based experiment. It could be that the conditions stipulated in the lab are likely to be different from those simulated in the real world. However, as we tried to replicate the dynamics of cyberattacks in the DG game, i.e., search followed by an attack, some of the conclusions derived from our experiment are likely to be valid for the real world. Furthermore, the size of the networks chosen across different conditions in the experiment was done to investigate the effect of increasing the number of web servers. However, these network sizes are likely to be different from those encountered in the real world. There may be some networks where the number of web servers is in the range as those chosen by us in the experiment. For such networks, some of the conclusions in this study may be useful. Finally, motivated by the real world, we assumed that adversaries did not possess knowledge about what web servers were honeypots and whether deception was present in a particular round. If the presence of deception and honeypots is known to adversaries, then it is likely that adversaries may take advantage of this knowledge and end up attacking a larger proportion of regular web servers.

Currently, we investigated the influence of network size in DG, where the proportion of honeypots was kept constant in the game. Another possibility is to vary the proportion of honeypots in the game with different network sizes and evaluate the combined influence of these variations on adversarial probe and attack actions. A second possibility is to test how the variation in the cost of probes and attack actions influences these actions. Still, a third possibility is to test a team of adversaries playing in networks of different sizes and with different proportions of honeypots. Some of these ideas form the immediate next steps in our program on behavioral cybersecurity.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Ethics Committee, Indian Institute of Technology, Mandi. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

HK contributed to the design of the game, implementation of experimental protocols, and data collection. ZM contributed to the data analyses and development of models. VD and PA developed the idea of the study, and contributed to the design, implementation of the study, and writing of the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the Department of Science and Technology (DST), Government of India, under the grant to VD and Dr. V. S. Chandrasekhar Pammi (IITM/DST-ICPS/VD/251). In addition, we are also grateful to the Indian Institute of Technology Mandi for providing the required resource for this project.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2020.535803/full#supplementary-material

TABLE S1 | Deception and non-deception rounds of deception game and instruction of deception game.

TABLE S2 | Data collected for analysis.

REFERENCES

Achleitner, S., La Porta, T. F., McDaniel, P., Sugrim, S., Krishnamurthy, S. V., and Chadha, R. (2017). Deceiving network reconnaissance using SDN-based virtual topologies. *IEEE Trans. Netw. Serv. Manag.* 14, 1098–1112. doi: 10.1109/tnsm.2017.2724239

Aggarwal, P., and Dutt, V. (2020). The role of information about an opponent's actions and intrusion detection alerts on cyber decisions in cyber security games. *Cyber Security* 3, 363–378.

Aggarwal, P., Gautam, A., Agarwal, V., Gonzalez, C., and Dutt, V. (2019). "Hackit: a human-in-the-loop simulation tool for realistic cyber deception experiments," in *Proceedings of the International Conference on Applied Human*

Factors and Ergonomics, (Cham: Springer), 109–121. doi: 10.1007/978-3-030-20488-4_11

Aggarwal, P., Gonzalez, C., and Dutt, V. (2016a). "Cyber-security: role of deception in cyber-attack detection," in *Advances in Human Factors in Cybersecurity*, ed. D. Nicholson (Cham: Springer), 85–96. doi: 10.1007/978-3-319-41932-9_8

Aggarwal, P., Gonzalez, C., and Dutt, V. (2016b). "Looking from the hacker's perspective: role of deceptive strategies in cyber security," in *Proceedings of the 2016 International Conference On Cyber Situational Awareness, Data Analytics And Assessment (CyberSA)*, (Piscataway, NJ: IEEE), 1–6.

Aggarwal, P., Gonzalez, C., and Dutt, V. (2017). "Modeling the effects of amount and timing of deception in simulated network scenarios," in *Proceedings of the 2017 International Conference On Cyber Situational Awareness, Data Analytics And Assessment (Cyber SA)*, (Piscataway, NJ: IEEE), 1–7.

Aggarwal, P., Gonzalez, C., and Dutt, V. (2020). "HackIt: a real-time simulation tool for studying real-world cyberattacks in the laboratory," in *Handbook of Computer Networks and Cyber Security*, eds D. Agrawal, G. M. Perez, D. Gupta, and B. B. Gupta (Cham: Springer), 949–959.doi: 10.1007/978-3-030-22277-2_39

Aggarwal, P., Moisan, F., Gonzalez, C., and Dutt, V. (2018). Understanding cyber situational awareness in a cyber security game involving recommendations. *Int. J. Cyber Situat. Awareness* 3, 11–38. doi: 10.22619/ijcsa.2018.100118

Almeshekah, M. H., and Spafford, E. H. (2016). "Cyber security deception," in *Cyber Deception*, eds S. Jajodia, V. Subrahmanian, V. Swarup, and C. Wang (Cham: Springer), 23–50. doi: 10.1007/978-3-319-32699-3_2

Bace, R., and Mell, P. (2001). *Special Publication on Intrusion Detection System*. Technical Report SP-800-31. Gaithersburg, MD: National Institute of Standards and Technology.

Bagchi, K. K., and Tang, Z. (2004). Network size, deterrence effects and Internet attack incident growth. *J. Inform. Technol. Theory Appl.* 6:9.

Camerer, C. F. (2003). *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton, NJ: Princeton University Press.

Cohen, F. (2006). The use of deception techniques: honeypots and decoys. *Handb. Inform. Security* 3, 646–655.

Cranford, E. A., Lebiere, C., Gonzalez, C., Cooney, S., Vayanos, P., and Tambe, M. (2018). "Learning about cyber deception through simulations: predictions of human decision making with deceptive signals in stackelberg security games," in *Proceedings of the 40th Annual Conference of the Cognitive Science Society*, At Madison, WI.

Dutt, V., Ahn, Y. S., and Gonzalez, C. (2013). Cyber situation awareness: modeling detection of cyber attacks with instance-based learning theory. *Hum. Fact.* 55, 605–618. doi: 10.1177/0018720812464045

Dutt, V., and Gonzalez, C. (2012). Making instance-based learning theory usable and understandable: the instance-based learning tool. *Comput. Hum. Behav.* 28, 1227–1240. doi: 10.1016/j.chb.2012.02.006

Dutt, V., Moisan, F., and Gonzalez, C. (2016). "Role of intrusion-detection systems in cyber-attack detection," in *Advances in Human Factors in Cybersecurity*, ed. D. Nicholson (Cham: Springer), 97–109. doi: 10.1007/978-3-319-41932-9_9

Field, A. (2013). *Discovering Statistics using IBM SPSS Statistics*. Thousand Oaks, CA: Sage.

Garcia-Teodoro, P., Diaz-Verdejo, J., Maciá-Fernández, G., and Vázquez, E. (2009). Anomaly-based network intrusion detection: techniques, systems and challenges. *Comput. Security* 28, 18–28. doi: 10.1016/j.cose.2008.08.003

Garg, N., and Grosu, D. (2007). "Deception in honeynets: a game-theoretic analysis," in *Proceedings of the 2007 IEEE SMC Information Assurance and Security Workshop*, (Piscataway, NJ: IEEE), 107–113.

Gonzalez, C., and Dutt, V. (2011). Instance-based learning: integrating sampling and repeated decisions from experience. *Psychol. Rev.* 118:523. doi: 10.1037/a0024558

Gonzalez, C., and Dutt, V. (2012). Refuting data aggregation arguments and how the instance-based learning model stands criticism: a reply to Hills and Hertwig (2012). *Psychol. Rev.* 119, 893–898. doi: 10.1037/a0029445

Gonzalez, C., Lerch, J. F., and Lebiere, C. (2003). Instance-based learning in dynamic decision making. *Cogn. Sci.* 27, 591–635. doi: 10.1207/s15516709cog2704_2

Heckman, K. E., Walsh, M. J., Stech, F. J., O'boyle, T. A., DiCato, S. R., and Herber, A. F. (2013). Active cyber defense with denial and deception: a cyber-wargame experiment. *Comput. Security* 37, 72–77. doi: 10.1016/j.cose.2013.03.015

Hope, C. (2020). *Cyber Losses Snowballing Despite an Increase in Cyber Security Spending*. Colorado Springs: Cyber Security.

Kiekintveld, C., Lisý, V., and Píbil, R. (2015). "Game-theoretic foundations for the strategic use of honeypots in network security," in *Cyber Warfare*, eds S. Jajodia, P. Shakarian, V. Subrahmanian, V. Swarup, and C. Wang (Cham: Springer), 81–101. doi: 10.1007/978-3-319-14039-1_5

La, Q. D., Quek, T. Q., Lee, J., Jin, S., and Zhu, H. (2016). Deceptive attack and defense game in honeypot-enabled networks for the internet of things. *IEEE Internet Things J.* 3, 1025–1035. doi: 10.1109/jiot.2016.2547994

Lenin, A., Willemson, J., and Sari, D. P. (2014). "Attacker profiling in quantitative security assessment based on attack trees," in *Proceedings of the Nordic Conference on Secure IT Systems*, (Cham: Springer), 199–212. doi: 10.1007/978-3-319-11599-3_12

Mason, W., and Suri, S. (2012). Conducting behavioral research on Amazon's Mechanical Turk. *Behav. Res. Methods* 44, 1–23. doi: 10.3758/s13428-011-0124-6

Matthews, K. (2019). *5 Futuristic Ways To Fight Cyber-Attacks*. Cologny: World Economic Forum.

Mell, P., Hu, V., Lippmann, R., Haines, J., and Zissman, M. (2003). *An Overview of Issues in Testing Intrusion Detection Systems*. Technical Report NIST IR 7007. Gaithersburg: National Institute of Standard and Technology.

PosTech (2020). *Attacks on Web Applications: 2018 in Review*. Available online at: https://www.ptsecurity.com/ww-en/analytics/web-application-attacks-2019/ (accessed July 10, 2020).

Rowe, N. C., and Custy, E. J. (2007). "Deception in cyber attacks," in *Cyber Warfare and Cyber Terrorism*, eds L. Janczewski and A. Colarik (Pennsylvania: IGI Global), 91–96. doi: 10.4018/978-1-59140-991-5.ch012

Sayegh, E. (2020). *More cloud, More Hacks: 2020 Cyber Threats*. Available online at: https://www.forbes.com/sites/emilsayegh/2020/02/12/more-cloud-more-hacks-pt-2/#1868047e69b3 (accessed February 15, 2020).

Shang, Y. (2018a). False positive and false negative effects on network attacks. *J. Stat. Phys.* 170, 141–164. doi: 10.1007/s10955-017-1923-7

Shang, Y. (2018b). Hybrid consensus for averager–copier–voter networks with non-rational agents. *Chaos Solitons Fractals* 110, 244–251. doi: 10.1016/j.chaos.2018.03.037

Shang, Y. (2019). Consensus of hybrid multi-agent systems with malicious nodes. *IEEE Trans. Circ. Syst. II Express Briefs* 67, 685–689. doi: 10.1109/tcsii.2019.2918752

Shimeall, T., and Spring, J. (2013). *Introduction to Information Security: A Strategic-Based Approach*. London: Newnes.

Symantec (2019). *Symantec Internet Security Threat Report 2019*. Available online at: https://img03.en25.com/Web/Symantec/%7Bdfc1cc41-2049-4a71-8bd8-12141bea65fd%7D_ISTR_24_2019_en.pdf (accessed February 14, 2020).

Trustwave (2019). *Trustwave Global Security Report 2019*. Available online at: https://www.trustwave.com/en-us/resources/library/documents/2019-trustwave-global-security-report (accessed February 15, 2020).

Wang, L., Jajodia, S., Singhal, A., and Noel, S. (2010). "k-zero day safety: measuring the security risk of networks against unknown attacks," in *Proceedings of the European Symposium on Research in Computer Security*, (Berlin: Springer), 573–587. doi: 10.1007/978-3-642-15497-3_35

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

REVIEW
published: 30 September 2020
doi: 10.3389/fpsyg.2020.01755

Human Cognition Through the Lens of Social Engineering Cyberattacks

Rosana Montañez[1], *Edward Golob*[2] *and Shouhuai Xu*[1]*

[1] Department of Computer Science, University of Texas at San Antonio, San Antonio, TX, United States, [2] Department of Psychology, University of Texas at San Antonio, San Antonio, TX, United States

Social engineering cyberattacks are a major threat because they often prelude sophisticated and devastating cyberattacks. Social engineering cyberattacks are a kind of psychological attack that exploits weaknesses in human cognitive functions. Adequate defense against social engineering cyberattacks requires a deeper understanding of what aspects of human cognition are exploited by these cyberattacks, why humans are susceptible to these cyberattacks, and how we can minimize or at least mitigate their damage. These questions have received some amount of attention, but the state-of-the-art understanding is superficial and scattered in the literature. In this paper, we review human cognition through the lens of social engineering cyberattacks. Then, we propose an extended framework of human cognitive functions to accommodate social engineering cyberattacks. We cast existing studies on various aspects of social engineering cyberattacks into the extended framework, while drawing a number of insights that represent the current understanding and shed light on future research directions. The extended framework might inspire future research endeavor toward a new sub-field that can be called *Cybersecurity Cognitive Psychology*, which tailors or adapts principles of Cognitive Psychology to the cybersecurity domain while embracing new notions and concepts that are unique to the cybersecurity domain.

Keywords: social engineering cyberattacks, cyberattacks, cyberdefenses, human cognition, cognitive psychology, cybersecurity

OPEN ACCESS

Edited by:
Paul Watters,
Independent Researcher, Melbourne, Australia

Reviewed by:
Alex Ng,
La Trobe University, Australia
Luca Allodi,
Eindhoven University of Technology, Netherlands

****Correspondence:***
Shouhuai Xu
shouhuai.xu@utsa.edu

Specialty section:
This article was submitted to Cognition, a section of the journal Frontiers in Psychology

Received: *19 January 2020*
Accepted: *25 June 2020*
Published: *30 September 2020*

Citation:
Montañez R, Golob E and Xu S (2020) Human Cognition Through the Lens of Social Engineering Cyberattacks. Front. Psychol. 11:1755. doi: 10.3389/fpsyg.2020.01755

1. INTRODUCTION

Social engineering cyberattacks are a kind of psychological attack that attempts to persuade an individual (i.e., victim) to act as intended by an attacker (Mitnick and Simon, 2003; Anderson, 2008). These attacks exploit weaknesses in human interactions and behavioral/cultural constructs (Indrajit, 2017) and occur in many forms, including phishing, scam, frauds, spams, spear phishing, and social media sock puppets (Stajano and Wilson, 2009; Linvill et al., 2019). For example, in the 2016 U.S. election, attackers used so-called social media sock puppets (also known as Russian Trolls) or fictional identities to influence others' opinions (Linvill et al., 2019). The effectiveness of current security technologies has made social engineering attacks the gateway to exploiting cyber systems. Most sophisticated and devastating cyberattacks often start with social engineering cyberattacks, such as spear phishing, where the attacker gains access into an enterprise network (Hutchins et al., 2011). Indeed, Mitnick and Simon (2003) describe many ways to gain access to secure systems using social engineering cyberattacks. Research in social engineering has mostly focused on understanding and/or detecting the attacks from a technological perspective (e.g., detecting phishing emails by analyzing email contents). However, there is no systematic understanding of the psychological components of these attacks, which perhaps explains why these attacks are so prevalent and successful . This motivates the present study, which aims to systematize human cognition through the lens of social engineering cyberattacks. To the best of our knowledge, this is the first of its kind in filling this void.

1.1. Our Contributions

In this paper, we make the following contributions. First, we advocate treating social engineering cyberattacks as a particular kind of psychological attack. This new perspective may be of independent value, even from a psychological point of view, because it lays a foundation for a field that may be called Cybersecurity Cognitive Psychology, which extends and adapts principles of cognitive psychology to satisfy cybersecurity's needs while embracing new notions and concepts that may be unique to the cybersecurity domain. This approach would pave the way for designing effective defenses against social engineering cyberattacks and assuring that they are built based on psychologically valid assumptions. For example, it may be convenient to assume that individuals are willing to participate in defenses against social engineering cyberattacks or that victims are simply reckless. However, these assumptions are questionable because most social engineering cyberattacks are crafted to trigger subconscious, automatic responses from victims while disguising these attacks as legitimate requests.

Second, as a first step toward the ultimate Cybersecurity Cognitive Psychology, we propose extending the standard framework of human cognition to accommodate social engineering cyberattacks. This framework can accommodate the literature studying various aspects of social engineering cyberattacks. In particular, the framework leads to a quantitative representation for mathematically characterizing *persuasion*, which is a core concept in the emerging Cybersecurity Cognitive Psychology and is key for understanding *behavior* in the traditional framework of human cognition. Some of our findings are highlighted as follows: (i) a high cognitive workload, a high degree of stress, a low degree of attentional vigilance, a lack of domain knowledge, and/or a lack of past experience makes one more susceptible to social engineering cyberattacks; (ii) awareness or gender alone does not necessarily reduce one's susceptibility to social engineering cyberattacks; (iii) cultural background does affect one's susceptibility to social engineering cyberattacks; (iv) the more infrequent the social engineering cyberattacks, the higher susceptibility to these attacks; (v) for training to be effective, it should capitalize on high-capacity unconscious processing with the goal of creating a warning system that operates in parallel with the user's conscious focus of attention; (vi) it is currently not clear how personality affects one's susceptibility to social engineering cyberattacks; and (vii) more studies, especially quantitative studies, need to be conducted to draw better and consistent results. In addition to these findings, we propose a range of future research directions, with emphasis on quantifying the effect of model parameters (i.e., victim's short-term cognition factors, long-term cognition factors, long-memory, and attacker effort) on the amount of persuasion experienced by the human target.

1.2. Related Work

To the best of our knowledge, we are the first to systematically explore the psychological foundation of social engineering cyberattacks. As discussed in the main body of the present paper, most prior studies focus on social engineering cyberattack or cyberdefense techniques. For example, Gupta et al. (2016) investigate defenses against phishing attacks; Abass (2018) discusses social engineering cyberattacks and non-technical defenses against them. Few prior studies have an aim that is similar to ours. Salahdine and Kaabouch (2019) review social engineering cyberattacks and mitigation strategies, but they do not discuss factors such as human cognition. Darwish et al. (2012) discuss at a high-level the relationship between human factors, social engineering cyberattacks, and cyberdefenses, but they neither examine what makes an individual susceptible to social engineering cyberattacks nor do they discuss the effect of a victim's psychological and situational conditions (e.g., culture and short-term factors) on the outcome. Pfleeger and Caputo (2012) take a multidisciplinary approach to examine cybersecurity from a Behavioral Science perspective, but they do not offer any systematic framework of looking at human cognition in the context of of social engineering cyberattacks. The lack of studies in social engineering cyberattacks might be associated with these studies involving human subjects. In an academic setting, approval for deceptive studies on human subjects requires consent from all entities involved, including ethics board and IT department (Vishwanath et al., 2011). The nature of the topic might also raise sensitivities among those involved (Jagatic et al., 2007), which can lengthen the process. This can be discouraging for most researchers.

1.3. Paper Outline

Section 2 reviews a basic framework for human cognition (without the presence of social engineering cyberattacks). Section 3 extends the basic framework to accommodate social engineering cyberattacks and systematizes victim's cognition through the lens of social engineering cyberattacks, with future research directions. Section 4 concludes the present paper.

2. OVERVIEW OF HUMAN COGNITION

In this section, we review human cognition functions prior to the presence of social engineering cyberattacks. This framework of human cognition serves as a basis for exploring how victims' cognition functions are exploited to wage social engineering cyberattacks.

2.1. Human Cognitive Functions

The term "cognition" can have radically different meanings in different contexts. Here, we use the term "cognition" in the broadest sense as a descriptive term for the software counterpart to the brain as hardware. That is, cognition is the abstract information processing that is being implemented by neurons in the brain (Pinker, 2009). From this perspective, cognition can also include information processing that computes emotions as well as the vast majority of neural information processing that is not reflected in our conscious awareness (Baars, 1997).

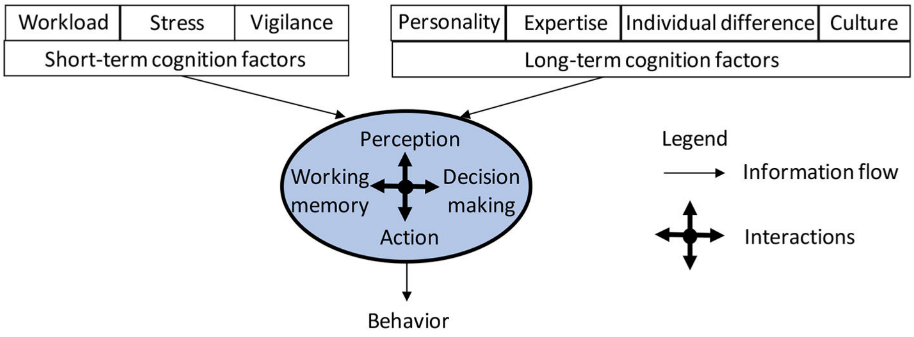


FIGURE 1 | A basic, selective schema of human cognition, where the blue background within the oval indicates long-term memory that is accessible by the four components of human cognition functions.

Cognitive psychologists often consider information processing to be the basic function of the brain, in the same way that the liver functions as a complex filter and arteries and veins are essentially pipes. Correlates of information processing in the brain can be directly observed using various methods to record electrical and chemical activity (Kandel et al., 2000). Information processing is evident at multiple spatial, from compartments within individual neurons to tightly organized networks having members in different parts of the brain. These concrete, physically measurable, neurophysiological activities are analogous to the hardware of a computer. Indeed, neurons have been profitably studied in terms of functioning as Boolean logic gates, and action potentials, perhaps the most characteristic property of neurons, is convey all-or-none binary states (Shepherd, 2004).

Figure 1 presents a very basic, and selective, schematic of human cognition functions, which are centered at four information processing components analogous to software components in an information processing system. These four components are called *perception*, *working memory*, *decision making*, and *action*. These four components are elaborated below as follows. Perception converts information in the word, sampled from the senses, into neural codes that can be used for intelligent behavior and conscious experience (Mesulam, 1998). Working memory consists of attention and short-term memory, and coordinates processing information by prioritizing certain information for short periods of time, often to accomplish a goal (Miyake and Shah, 1999). Decision making further prioritizes information from working memory and other unconscious sources and is a gateway to behavior (Kahneman, 2011). Action is the implementation of computations from decision making, as well as other influences, and also organizes the physical activity of muscles and glands that are measurable as behavior (Franklin and Wolpert, 2011). Perception, working memory, decision making, and action are often considered to be roughly sequential, as when trying to hit a baseball, but can mutually influence each other in many ways. All of these cognitive processes operate on a foundation of accumulated knowledge in memory, which informs these processes, such as when perceiving a familiar face.

Memory is intrinsic to cognition, because information processing occurs over time and thus requires some information to be retained over time. The basic processes of perception, working memory, decision making, and action that are engaged "in the moment" use information that is preserved from earlier moments in time. Memory consists of distinct systems (Tulving and Craik, 2000), in the same way that our domain of "perception" includes the visual, auditory, somatosensory, olfactory, gustatory, and vestibular systems. One important distinction among systems is whether the information is retained over short periods of time, typically seconds to minutes, or longer periods of time. In our overview shown in **Figure 1**, short-term memory is a component of working memory. Long-term memory contributes to cognition in general, and for this reason, we have situated all of four domains supporting cognition in the moment within long-term memory (indicated by the blue background). As with the other cognitive domains, memory systems can work in parallel. For example, the memory of the previous sentence is supported by short-term memory, yet the memory for what each word means resides in long-term memory.

Above, we presented several basic types of information processing that together generate behavior. We now consider how these basic cognitive processes can be influenced, for better or worse, by a few important factors that are demonstrably relevant to cybersecurity. The "short-term" factors, reflecting the immediate situation, and "long-term" factors are ultimately coded in some form by the brain and exert an influence on the basic cognitive processes that drive behavior. The short-term and long-term factors are elaborated in the next two subsections.

2.2. Short-Term Cognitive Factors

We focus on three short-term factors: *workload*, *stress*, and *vigilance*. These factors operate on relatively short timescales (minutes to hours) that have been intensively studied because they impair human performance. We will consider how these factors may relate to social engineering and point out the extant literature and promising future directions.

2.2.1. Workload

Human cognition is affected by cognitive workload, which depends on task demand and the operator in question. Depending on the details, two tasks can be done at the same time with little or no performance costs (a manageable workload) or be nearly impossible to do well together (a very high workload). A nice example comes from Shaffer (1975), who found that typists could very accurately read and type while they also verbally repeated a spoken message. Performance, however, plummeted on both tasks if they tried to take dictation (typing a spoken message) while also trying to read a written message out loud. The differences are thought to reflect the use of phonological (sound-based) and orthographic (visual letter-based) cognitive codes. In the first example one code is used per task (phonological: listen to speech-talk; orthographic: read-type), while in the second each code is used for both tasks (speech-type; read-talk). To account for these complexities, psychologists have developed theories that consider different types of cognitive codes (Navon and Gopher, 1979), such as auditory or visual sensory input, higher-level verbal or spatial codes, and output codes for driving speech or manual behaviors (Wickens, 2008). Measures have also been developed to quantify the subjective sense of how much "cognitive work" is being done in a given task. Perhaps the most common instrument to measure subjective workload is the NASA-TLX, which has six dimensions that are clearly explained to the subject, such as "mental demand" or temporal demand (time pressure), and are rated on a scale from low to high. Lastly, neurophysiological measures are often used to provide objective, convergent measures of workload as well as suggest potential neural mechanisms. Neurophysiological measures such as transcranial Doppler measures of blood flow velocity in the brain, EEG measures of brain electrical potentials, autonomic nervous system activity such as skin conductance and heart rate and its variability, and functional magnetic resonance imaging (MRI) to quantify changes in blood flow that are secondary to neural activity are commonly used (Parasuraman and Rizzo, 2008).

2.2.2. Stress

Acute stress may also influence cognition and behaviors that are relevant to cybersecurity. We distinguish acute from chronic stress, with chronic stress beginning after a duration on the order of months, as their impact on cognition can differ and chronic stress is better classified here as a long-term factor. The neurobiological and hormonal responses to a stressful event have been well-studied, as have their impact on behavior (Lupien et al., 2009). Acute stress can influence attention, a vital component to working memory, in ways that are beneficial as well as detrimental (Al'Absi et al., 2002). Attentional tunneling is one such effect of acute stress where attention is hyper-focused on aspects relevant to the cause of the stress but is less sensitive to other information. The term tunneling derives from the use of spatial attention tasks, where arousal due to stress leads to subjects ignoring things that are more distant from the focus of attention (Mather and Sutherland, 2011). In the realm of cyber security, attention tunneling from an emotion-charged phishing message could lead one to hyper-focusing on the email text but ignore a suspicious address or warnings at the periphery. Working memory is also vulnerable to acute stress (Schwabe and Wolf, 2013), particularly by way of interfering with prefrontal cortex function (Elzinga and Roelofs, 2005; Arnsten, 2009). Decision making can be driven in two fundamentally different ways (Evans, 2008). The first is by relatively automatic processes that are fast but may not be the optimal choice in some instances (termed "heuristics" and "biases") (Tversky and Kahneman, 1974; Gigerenzer, 2008). The second approach is by using conscious, controlled processing reasoning, which is slower but can be more sensitive to the particulars of a given situation. Acute stress has a variety of effects on decision making and many subtleties (Starcke and Brand, 2012), but it can, in general, impair rational decision making, and one way is by reducing the likelihood of controlled decision making and increasing the use of automatic processing.

2.2.3. Vigilance

Vigilance and sustained attention are two closely related, sometimes synonymous, terms for the concept that cognitive performance will systematically change the longer you perform a given task. Here we will use the term vigilance, which in the laboratory is studied in sessions that typically last 30–60 min. In a classic work by Mackworth (1948), subject watched an analog clock and responded to intermittent jumps in the clock hand. Much work since then has showed that performance in a wide range of tasks declines substantially over these relatively short periods of time (termed the "vigilance decrement") (Parasuraman and Rizzo, 2008). In our view, the potential impact of the vigilance decrement on behavior is an important factor to explore, because the probability of user error may covary with time on task. For example, the likelihood of downloading malware may increase as users go through their email inbox, particularly if they have limited time. Lastly, we note that although the situational categories of workload, stress, and vigilance are individually important to examine in the realm of cybersecurity, they are also known to interact with each other. For example, a high workload and prolonged vigilance are stressful (Parasuraman and Rizzo, 2008). Another distinction to keep in mind is that many laboratory vigilance tasks are boring and have a low workload. The extent that the vigilance literature generalizes to other settings such as an office, where workers may have high workloads and stress from complex job demands, is an empirical question worth considering in future cybersecurity studies.

2.3. Long Term Cognitive Factors

In contrast to short-term factors that reflect the current situation and can change rapidly, our second grouping of "long-term factors" covers more stable attributes of a person and their experiences that only gradually change. We consider factors of personality, expertise, age and gender, and culture. We include personality as a long-term factor, even though it can be situation dependent as well (as with short-term factors) (Kenrick and Funder, 1988). These factors offer some predictability of individual behavior in a given situation. In the context of cybersecurity, long-term psychological factors can impact how an individual responds to social engineering attacks.

2.3.1. Personality

To Psychologists, "personality" is a technical term that differs somewhat from ordinary usage. It refers to individual differences in thoughts, feelings, and behaviors that are relatively consistent over time and situations. We say "relatively" because, as noted above, thoughts, feelings, and behaviors are highly dependent on the situation, and lifespan approaches have defined notable changes in personality with age (Donnellan and Robins, 2009). Personality research is dominated by the Big 5 framework of personality domains, which was developed over much of the twentieth century in various forms (Digman, 1997). The Big 5 framework is based on statistical methods (factor analysis) that identify abstract dimensions that can economically account for much of the variance in personality measures. The factors are labeled conscientiousness, agreeableness, neuroticism, openness to experience, and extraversion. For present purposes, the labels of the factors are adequate descriptions of the underlying constructs. Many studies on the relationship between social engineering and personality focus on openness, conscientiousness, and neuroticism which are thought to have the most impact on susceptibility to social engineering. The factors that comprise the Big 5 framework are the following:

1. Openness: the willingness to experience new things.
2. Conscientiousness: favors-norms, exhibiting self-control and self-discipline, and competence.
3. Extraversion: being more friendly, outgoing, and interactive with more people.
4. Agreeableness: being cooperative, eager to help others, and believing in reciprocity.
5. Neuroticism: tendency to experience negative feelings, guilt, disgust, anger, fear, and sadness.

2.3.2. Expertise

Expertise is typically limited to relatively narrow domain and does not transfer to other areas as much as we tend to believe (termed the "transfer problem") (Kimball and Holyoak, 2000). Limited transfer of expertise can be compounded by cognitive illusions such as the Dunning-Kruger effect. The Dunning-Kruger effect empirically shows that individuals often overestimate their competence relative to their objective performance (Kruger and Dunning, 1999). Similarly, the "illusion of knowledge" shows that people generally know far less about a topic than they believe, as revealed by questioning (Keil, 2003). In the realm of cybersecurity, these and other empirical phenomena underpin user over confidence. As will be detailed below, narrow expertise about cybersecurity can be beneficial, but computer expertise more generally may not confer security benefits.

2.3.3. Individual Differences

There are many kinds of individual differences and we focus on two kinds: age and sex/gender; others would include role in companies and seniority. In terms of age, there are dramatic changes in cognitive function and behavioral capacities of children as they develop (Damon et al., 2006). Considering how youths can safely use computers is a major parenting, education, public policy, and law enforcement challenge. Social engineering attacks can readily take advantage of the cognitive and emotional vulnerabilities of children, and countermeasures are often quite different than with adults (see below). Cognition changes throughout the adult lifespan at a less frenetic pace vs. in children, but longer-term changes are similarly dramatic (Park and Reuter-Lorenz, 2009; Salthouse, 2012). Declines in fluid intelligence, essentially one's ability to "think on your feet," are particularly dramatic and have wide implications for everyday life (Horn and Cattell, 1967). Overall, there are many changes, some declining with age (fluid intelligence) and others not (Schaie, 2005). Another angle is that age is positively associated with the risk for many neurological disorders that can impair cognition, such as stroke and Alzheimer's disease (Hof and Mobbs, 2001). Age-related neurological disorders are not considered "normal aging," but the potential vulnerability of many elders due to brain disease has been well-known to criminals for a long time. As expected, social engineering attacks are a major problem for this vulnerable population.

Psychology has a long history of studying sex differences, defined by biology (i.e., the presence of two X or one X and one Y chromosome) and gender, which is a social, rather than biological, construct. In terms of basic cognitive functions such as working memory and decision making, which are typically studied in a neutral laboratory context (such as remembering strings of letters, judging categories of pictures, etc.) there are generally little or no differences between sexes and genders. There are a few well-documented exceptions, such as males having an advantage for mental spatial rotations (Voyer et al., 1995). The situation is quite different when examining cognition in the context of social and emotional factors (Cahill, 2006). For our purposes, sex and gender are basic considerations for social engineering attacks, particularly spear phishing, which is tailored to an individual. Our list could include many other types of individual differences that are useful for social engineering attacks, such as socio-economic class, education, personal interests, job position. We chose to focus on age and sex/gender because they are prominent topics in the cognition literature and important considerations for cyber security challenges such as spear phishing.

2.3.4. Culture

In mainstream cognitive psychology, culture is not a prominent variable, as much of the basic literature studies participants in countries that have predominantly western cultures (Arnett, 2008). Nonetheless, a wide variety of studies have shown that cultural differences are evident in many aspects of cognition, such as basic perception, language and thought, attention, and reasoning (Grandstrand, 2013). Culture is an important variable to consider for any social engineering attack. A phishing email, for example, is unlikely to be effective if the message violates norms of the target's culture. We also consider the more specific case of organizational culture in the workplace because it is highly relevant to employee behavior as it applies to cyber security (Bullee et al., 2017). As with all of the other short-and long-term variables that we consider, culture is assumed to interact with the other variables, with particularly large interactions with age, gender, and perhaps personality.

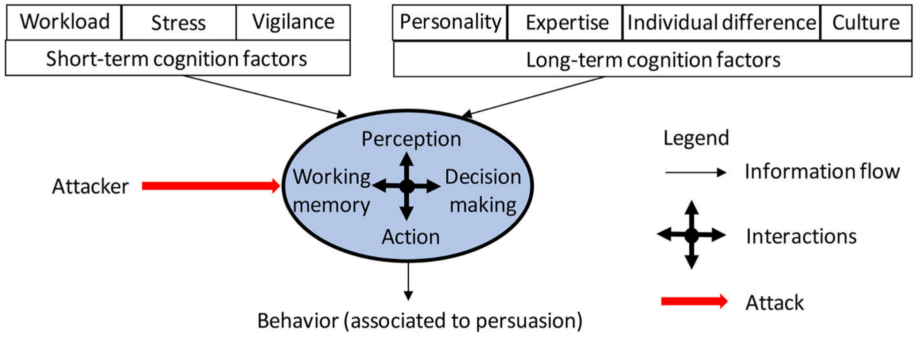


FIGURE 2 | Extending the basic schema of human cognition presented in **Figure 1** to accommodate social engineering cyberattacks. The extension is to incorporate an attacker that wages a social engineering cyberattack against a victim's human cognition functions (i.e., the oval). The resulting behavior is associated to persuasion (i.e., an attack succeeds when a victim is persuaded to act as intended by the attacker).

3. VICTIM COGNITION THROUGH THE LENS OF SOCIAL ENGINEERING CYBERATTACKS

Social engineering cyberattacks are a type of psychological attack that exploits human cognition functions to persuade an individual (i.e., victim) to comply with an attacker's request (Anderson, 2008). These attacks are centered around a social engineering message crafted by an attacker with the intent of persuading a victim to act as desired by the attacker. These attacks often leverage behavioral and cultural constructs to manipulate a victim into making a decision based on satisfaction (gratification), rather than based on the best result (optimization) (Kahneman, 2011; Indrajit, 2017). For example, one behavioral construct is that most individuals would trade privacy for convenience, or bargain release of information for a reward (Acquisti and Grossklags, 2005).

To establish a systematic understanding of the victim's cognition through the lens of social engineering cyberattacks, we propose extending the framework presented in **Figure 1** to accommodate social engineering cyberattacks against human victims' cognition functions, leading to the framework highlighted in **Figure 2**. This implies that the resulting behavior of a victim will also depend on the attacker's effort. In what follows, we will cast the social engineering cyberattacks literature into this framework, by first discussing the literature related to short-term and long-term cognition factors, and then the literature related to cognition functions.

3.1. Short-Term Cognition Factors Through the Lens of Social Engineering Cyberattacks

3.1.1. Workload

In computer-mediated communications, cognitive workload can affect an individual's ability to process socially engineered messages. Pfleeger and Caputo (2012) observe that cognitive workload could make individuals overlook elements that are not associated with the primary task. This effect, called inattentional blindness, affects an individual's ability to notice unexpected events when focusing on the primary task (Simons, 2000). In most cases, security is a secondary task . For example, when an employee attempts to manage several tasks simultaneously (e.g., reply to hundreds of emails in the email inbox while answering calls and an occasional request from the boss), the employee is more likely to overlook cues in phishing messages that might indicate deception. A study that examined actual phishing behavior by sending employees an innocuous phishing email, found that self-perceived work overload was positively associated with the likelihood of clicking on the phishing link (Jalali et al., 2020). Vishwanath et al. (2011) investigate the effect of information processing and user's vulnerability to phishing. Leveraging two phishing attacks that target a university, they survey undergraduate students on their recollection and response to the phishing emails. They find that in the presence of a perceived relevant email, individuals focus more on urgency cues, while overlooking deception cues in the message, such as sender's email address or email grammar/spelling. They also find that individuals that regularly manage large volumes of emails have a high inattentiveness when evaluating emails, making them more vulnerable to phishing attacks. They also find that a high *email load* triggers an automatic response, meaning that workload significantly increases a victim's vulnerability to phishing attacks.

Summarizing the preceding discussion, we draw:

INSIGHT 1. *Cognitive workload, via mechanisms such as inattentional blindness, can increase vulnerability to social engineering cyberattacks.*

3.1.2. Stress

The particular kind of stress, namely acute stress mentioned above, has only been indirectly investigated in the context of social engineering cyberattacks. Stajano and Wilson (2009) examine how principles of scams apply to systems security. Scams are a form of social engineering cyberattack that usually involves a physical interaction between attacker and victim. One scamming technique is the principle of distraction, by which the attacker can take advantage of a victim that is in a state of mind that prevents them from evaluating deceptive cues. For example, when an unemployed individual pays a job recruiting company for job hunting assistance, the individual does not realize that it is a scam. Catphishing is a social engineering cyberattack by which the attacker creates a fictional online persona to lure a victim into a romantic relationship for financial gains. In this case, an individual who is searching for a romantic partner and is experiencing some personal stress might find a catphishing message appealing and, therefore, may be unable to detect the deception cues in the catphishing messages. In summary, we draw the following:

INSIGHT 2. *Stress may reduce one's ability to detect deception cues in social engineering cyberattack messages but the direct effects of acute stress on cybersecurity social engineering have not been examined.*

3.1.3. Vigilance

Purkait et al. (2014) conduct a study to examine cognitive and behavioral factors that affect user's capabilities in detecting phishing messages. They study attentional vigilance and short-term memory by surveying 621 participants' ability to identify phishing sites, Internet skills, usage and safe practices, and demographics. The measure of "vigilance" was a brief visual search task in six photographs, which did not evaluate vigilance as we conventionally defined it above. Individual differences in these visual search scores were significant predictors of performance distinguishing spam from phishing websites, which likely reflects the ability to detect visual cues on the website that distinguish spam from phish sites.

INSIGHT 3. *Attentional vigilance, particularly the vigilance decrement, may be an important influence on susceptibility to social engineering attacks, but more research is needed.*

3.2. Long-Term Cognition Factors Through the Lens of Social Engineering Cyberattacks

3.2.1. Personality

Personality has been extensively studied in the context of phishing. Studies show that Big 5 personality traits are related to individuals' susceptibility to social engineering cyberattacks. Pattinson et al. (2012) study how personality traits and impulsiveness affect behavioral responses to phishing messages. They find that individuals that score high on extraversion and openness manage phishing emails better than individuals with other personality types. Halevi et al. (2013) find that high neuroticism increases responses to prize phishing messages and that individuals with a high openness have low security setting on social media account, increasing their exposures to privacy attacks. Halevi et al. (2016) find that personality traits affect security attitudes and behaviors as follows: high conscientiousness is associated to highly secure behaviors but does not affect self-efficacy (i.e., one's ability in independently resolving computer security issues); high openness is associated to high self-efficacy; high neuroticism is associated to low self-efficacy; and high emotional stability (inverse of neuroticism) is associated to high self-efficacy. Cho et al. (2016) contradict some of the findings presented in Halevi et al. (2013), by finding that high neuroticism decreases trust and increases risk perception, which makes one more likely to misclassify benign emails as phishing ones. They also find that higher agreeableness increases trust and lowers risk perception (i.e., more likely classifying phishing messages as benign). Consciousness is commonly associated with self-control, which diminishes impulsive behavior (Cho et al., 2016). Pattinson et al. (2012) find that less impulsive individuals manage phishing messages better. Halevi et al. (2015) show that individuals with high consciousness and lower risk perception are more likely to fall victims to social engineering cyberattack messages. Lawson et al. (2018) find that extroversion decreases phishing detection accuracy while high consciousness increases detection accuracy, and that openness is associated with higher accuracy in detecting legitimate messages. Darwish et al. (2012) find that individuals high in extraversion and agreeableness pose a higher security risk. McBride et al. (2012) find that consciousness is associated with low self-efficacy and threat severity. Workman (2008) and Lawson et al. (2018) that personality traits are related to the degree of persuasion by social engineering cyberattacks. Summarizing the preceding discussion, we draw the following:

INSIGHT 4. *Literature results are not conclusive on how personality may influence one's susceptibility to social engineering cyberattacks.*

3.2.2. Expertise

Related to expertise, domain knowledge, awareness, and experience have been studied in the literature on their impact on reducing one's susceptibility to social engineering cyberattacks.

Impact of domain knowledge. An individual's knowledge related to cyberattacks increases their capability to resist social engineering cyberattacks. For example, the knowledge can be about web browsers, including how to view site information and evaluate certificates. Kumaraguru et al. (2006) find (i) non-expert individuals consider fewer security indicators (e.g., meaningful signals) than experts; (ii) non-expert individuals used simple rules to determine the legitimacy of a request, while experts also consider other useful information (e.g., context) that may reveal security concerns with the request; (iii) non-expert individuals make decisions based on their emotions, while experts make their decisions based on reasoning; and (iv) non-expert individuals rely more on (spoofable) visual elements to make decisions because they lack the knowledge that security indicators can be compromised, while experts are more efficient at identifying suspicious elements in a message. For example, corresponding to (iii), they observe that a non-expert individual might decide to download a software program based on how much they *want* it and if the downloading website is recognizable; whereas an expert might consider how much they *need* it and if the downloading website is a reputable source. These findings resonate with what is found by Klein and Calderwood (1991), namely, that experts make decisions based on pattern recognition, rather than purely analyzing the available options. Byrne et al. (2016) find that risk perception for non-expert individuals is influenced by the benefit that can be obtained for an activity, meaning that actions that an individual considers beneficial are performed more often and are perceived as less risky.

INSIGHT 5. *Domain knowledge helps reduce vulnerability to social engineering cyberattacks.*

Impact of awareness. As a rule of thumb, training on non-expert individuals often emphasize on awareness. In a study of victims in frauds involving phishing and malware incidents, Jansen and Leukfeldt (2016) find that most participants express that they have knowledge of cybersecurity, but it turns out only a few of them indeed have the claimed knowledge. Downs et al. (2006) find that awareness of security cues in phishing messages does not translate into secure behaviors because most participants are unable to tie their actions to detrimental consequences. On the other hand, it may be intuitive that individuals that have received formal computer education would be less vulnerable to social engineering cyberattacks. To the contrary, Ovelgönne et al. (2017) find that software developers are involved in more cyberattack incidents when compared to others. Purkait et al. (2014) find that there is no relationship between one's ability to detect phishing sites and one's education and technical backgrounds, Internet skills, and hours spent online. Halevi et al. (2013), Junger et al. (2017), and Sheng et al. (2010) find that knowledge acquired through priming and warning does not affect ones' susceptibility to phishing attacks.

INSIGHT 6. *Awareness and general technical knowledge do not necessarily reduce one's susceptibility to social engineering cyberattacks, perhaps because human cognition functions have not been taken into consideration.*

Impact of experience. Harrison et al. (2016) find that knowledge about phishing attacks increases one's attention and elaboration when combined with subjective knowledge and experience, and therefore lowers one's susceptibility to fall victim to social engineering cyberattack messages. Wang et al. (2012) find that knowledge about phishing attacks increases one's attention to detect indicators. Pattinson et al. (2012) find that the higher familiarity with computers, the higher capability in coping with phishing messages. Wright and Marett (2010) find (i) a combination of knowledge and training is effective against phishing attacks; (ii) individuals with a lower self-efficacy (i.e., one's ability to manage unexpected events) and web experience are more likely to fall victims to social engineering cyberattacks; and (iii) individuals with high self-efficacy are less likely to comply with information requests presented in phishing attacks. Halevi et al. (2016) find that a high self-efficacy correlates a better capability to respond to security incidents. Arachchilage and Love (2014) find that self-efficacy, when combined with knowledge about phishing attacks, can lead to effective strategies for coping with phishing attacks. Wright and Marett (2010) find that experiential factors (e.g., self-efficacy, knowledge, and web experience) have a bigger effect on individuals' response to phishing attacks than dispositional factors (e.g., the disposition to trust and risk perception). Van Schaik et al. (2017) find that a higher risk perception of online threats is associated with exposure to the knowledge that is specific to the threat. Downs et al. (2006) find that users can detect social engineering cyberattacks that are similar to the ones they have been exposed to. Redmiles et al. (2018) find that the more time an individual spends online, the more skilled they are at identifying spams, and the less likely they will click on the links in the spam messages. Gavett et al. (2017) find that education and previous experience with phishing attacks increased suspicion on phishing sites. Cain et al. (2018) find that past security incidents do not significantly affect secure behaviors. Abbasi et al. (2016) find (i) older, educated females and males fell victim to phishing attacks in the past are less likely to fall victim to phishing attacks again; (ii) young females with low phishing awareness and previous experience in suffering from small losses caused by phishing attacks do not necessarily have a lower susceptibility to phishing attacks in the future; and (iii) young males with high self-efficacy and phishing awareness and previous experiences in phishing attacks also do not necessarily have a lower susceptibility to phishing attacks in the future.

INSIGHT 7. *Self-efficacy, knowledge, and previous encounter of social engineering cyberattacks collectively reduce one's susceptibility to social engineering cyberattacks. In particular, costly phishing experiences would greatly reduce one's susceptibility to social engineering cyberattacks, while non-costly experiences do not.*

3.2.3. Individual Differences

Two kinds of individual differences have been investigated in the context of social engineering cyberattacks: gender and age.

Impact of gender. Initial studies suggest a relationship between gender and phishing susceptibility. Hong et al. (2013) finds that individual differences (e.g., dispositional trust, personality, and gender) are associated with the ability to detect phishing emails. Halevi et al. (2015) find that for women, there is a positive correlation between conscientiousness and clicking on links and downloading files associated with phishing attacks. Halevi et al. (2013) find that women exhibit a strong correlation between neurotic personality traits and susceptibility to phishing attacks, but no correlation to any personality trait is found for men. Halevi et al. (2016) reports that women exhibit lower self-efficacy than men. Sheng et al. (2010) find that women with low technical knowledge are more likely to fall victim to phishing attacks. Sheng et al. (2010) find that women are more likely to fall victim to phishing attacks.

However, later studies provide a different view. Sawyer and Hancock (2018) finds that there is no relationship between gender and phishing detection accuracy. Similarly, Purkait et al. (2014) find that there is no relationship between gender and the ability to detect phishing sites. Byrne et al. (2016) finds that there is no relationship between gender and risk perception. Rocha Flores et al. (2014) finds that there is no significant correlation between phishing resiliency and gender. Bullee et al. (2017) finds that gender does not contribute to phishing message responses. Abbasi et al. (2016) finds (i) women with a high self-efficacy have a low susceptibility to social engineering cyberattacks, and that women without awareness of the social engineering cyberattack threat have a high susceptibility to these attacks; and (ii) men with previous costly experiences with phishing attacks have a low susceptibility to these attacks, while overconfidence increases the susceptibility to these attacks. Cain et al. (2018) find that although men may have more knowledge about cybersecurity than women, there is no difference in terms of insecure behaviors by gender. Redmiles et al. (2018) show that, in the context of social media spam, gender affects message appeal but not susceptibility to social engineering cyberattacks, and that women are more likely to click on sales-oriented spams while men are more likely to click on media spams that feature pornography and violence. Goel et al. (2017) find that women open more messages on prize reward and course registration than men. Rocha Flores et al. (2014) find that gender affects the type of phishing message an individual would respond to and that women are less susceptible than men to generic phishing messages.

INSIGHT 8. *Gender does not have a big impact on the susceptibility to social engineering cyberattacks.*

Impact of age. Most studies focus on age groups in young people (18–24) and old (45+) ones. In general, youth is related to inexperience, high emotional volatility (Zhou et al., 2016), less education, less exposure to information on social engineering, and fewer years of experience with the Internet. These factors are often accompanied by a low aversion to risk and therefore can increase the chances of falling victim to social engineering cyberattacks (Sheng et al., 2010). In an experiment involving 53 undergraduate students in the age group of 18–27, Hong et al. (2013) find that the students' confidence in their ability to detect phishing messages does not correlate to their detection rate. Sheng et al. (2010) investigate the relationship between demographics and susceptibility to phishing attacks and find that individuals at the age group of 18–25 are more susceptible to phishing attacks than other groups 25+. Lin et al. (2019) report a similar result but for an old group. Howe et al. (2012) find that age also affects risk perception: individuals in the age groups of 18–24 and 50–64 perceive themselves at lower security risks compared to other groups and therefore are more susceptible to social engineering cyberattacks. Purkait et al. (2014) find that the detection of phishing messages decreases with the age and frequency of online transactions. Bullee et al. (2017) find that age has no effect on spear-phishing responses and that Years of Service (YoS) is a better indicator of victimization (i.e., a greater YoS means less likely susceptible to social engineering cyberattacks). Gavett et al. (2017) examine the effect of age on phishing susceptibility and showed that processing speed and planning executive functions affect phishing suspicion, hinting a relationship between phishing susceptibility and cognitive degradation from aging.

INSIGHT 9. *Old people with higher education, higher awareness and higher exposure to social engineering cyberattacks are less susceptible to these attacks.*

3.2.4. Culture

Culture affects individuals' online activities (Sample et al., 2018), decision making process and uncertainty assessment (Chu et al., 1999), development of biases and risk perception (Pfleeger and Caputo, 2012; da Veiga and Martins, 2017), reactions to events (Hofstede et al., 2010; Rocha Flores et al., 2014), and self-efficiency (Sheng et al., 2010; Halevi et al., 2016). Redmiles et al. (2018) suggest that country/communal norms might affect spam consumption as follows: in countries where spam is more prevalent, users are 59% less likely to click on spam when compared to countries where spam is less prevalent. Halevi et al. (2016) find that individuals with higher risk perception have higher privacy attitudes, which reduce the susceptibility to social engineering cyberattacks. Al-Hamar et al. (2010) perform experimental spear-phishing attacks against two groups from Qatar, where one group consists of 129 employees of a company (dubbed *employees*) and the other of 30 personal acquaintances (dubbed *friends*); they find that find that 44% of the individuals in the *employees* group are successfully phished while 57% of the *friends* groups are successfully phished. Tembe et al. (2014) report that participants from India exhibit a higher susceptibility to phishing attacks when compared with participants from the USA and China. (Bullee et al., 2017) report that participants from China and India might not be aware of the harms and consequences of phishing attacks, while participants from the USA exhibit more awareness of privacy and online security features (i.e., SSL padlocks) and are more active in safeguarding their personal information. Halevi et al. (2016) find that although culture is a significant predictor of privacy attitude, it does not predict security behavior and self-efficacy. Bohm (2011) finds that culturally sound messages do not raise suspicion. Farhat (2017) and Hofstede et al. (2010) show that scams with culture-specific shame appeal are more likely to be effective in a certain culture. Bullee et al. (2017) find that participants from countries with a higher Power Distance Index (PDI), which means that individuals are more likely to comply with hierarchy, are more vulnerable to phishing than those individuals from countries with a lower PDI. Sharevski et al. (2019) show how to leverage cultural factors to tailor message appeal.

INSIGHT 10. *Culture affects privacy and trust attitudes, which indirectly affect one's susceptibility to social engineering cyberattacks.*

3.3. Victim Cognition Functions Through the Lens of Social Engineering Cyberattacks

3.3.1. Long-Term Memory

As reviewed in section 2.1, long-term memory is a very broad field, of which the following aspects have been studied through the lens of social engineering cyberattacks. The first aspect is the *frequency of attacks*. The environment in which a victim operates provides a context that may be exploited by an attacker. For example, the attacker may leverage an ongoing societal incident or personal information to craft messages to make the victim trust these messages. The attacker attempts to build trust with a victim while noting that a suspicion thwarts the attack (Vishwanath et al., 2018). Both trust and suspicion are affected by the environment, such as the frequency of the social engineering events exploited by the social engineered messages. For example, in a situation where social engineering cyberattacks are expected, the attacker is at a disadvantage (Redmiles et al., 2018); in a situation where social engineering cyberattacks are infrequent, the attacker has the advantage. Sawyer and Hancock (2018) investigate how infrequent occurrence of phishing events (i.e., the prevalence of phishing) affects individuals' abilities to detect cyberattacks delivered over emails. In their experiment, they ask three groups to identify malicious and legitimate email messages. The three groups, respectively, contain 20, 5, and 1% malicious emails. They find that the accuracy of the detection of malicious emails is lower for the group dealing with emails that contain 1% malicious ones. Similarly, Kaivanto (Kaivanto, 2014) show that a lower probability of phishing occurrence increases victim's susceptibility to phishing cyberattacks.

INSIGHT 11. *The success of social engineering cyberattacks is inversely related to their prevalence.*

Insight 11 causes a dilemma: when automated defenses are effective at detecting and filtering most social engineering cyberattacks, the remaining attacks that make it through to users are more likely to succeed. One approach to dealing with this dilemma is to resort to principles in Cognitive Psychology. It is known that most of the brain's information processing is sealed-off from conscious awareness (Nisbett and Wilson, 1977), some permanently while other information could be consciously appreciated, but may not be conscious at a given moment. Our visual system, for example, computes 3-D depth from 2-D retinal inputs (DeValois and DeValois, 1990). We do not consciously experience the calculations needed to transform the 2-D input into a 3-D percept. Instead, we are aware of the product (i.e., seeing a 3-D world) but not the process that led to the product. The influences of subconscious processing are well-known to impact behavior (Kahneman, 2011; Nosek et al., 2011). This fact leads to the following insight:

INSIGHT 12. *Training methods that ask people to consciously think about social engineering cyberattacks are unlikely to be very successful unless the learning reaches the point where it is a habit that, largely unconsciously, guides safer computer use behavior.*

Insight 12 would avoid the dilemma mentioned above because when the training/learning effort reaches the point that users can deal with social engineering cyberattacks subconsciously, users can effectively defend these attacks. This coincides with findings of Halevi et al. (2015), Rocha Flores et al. (2014), Halevi et al. (2016), Howe et al. (2012), and Sheng et al. (2010), namely that users with higher risk perception can reduce the chance they fall victim to social engineering cyberattacks.

3.3.2. Victim Cognition Functions: A Preliminary Mathematical Representation

The framework described in **Figure 2** formulates a way of thinking in modeling how the behavior of a victim is influenced by the victim's short-term cognition factors (or `short_term factors`), long-term cognition factors (or `long_term factors`) and long-memory (or `long_memory`) as well as the attacker's effort (or `attacker_effort`). This formulation is applicable to phishing, spear phishing, whaling, water holing, scams, angler phishing, and other kinds of social engineering attacks, where the resulting behavior is whether a victim is persuaded by the attacker to act as intended. For example, spear-phishing is a special case of the model because the attacker often makes a big effort at enhancing *message appeal* by exploiting personalization; scam is another special case of the model because the attacker often makes a big effort at enhancing *message appeal* by exploiting situational setting and possibly stress. In principle, the behavior (`behavior`) of social engineering cyberattacks can be described as some mathematical function f (mathematically speaking, more likely it will be a family of functions):

$$\begin{aligned}\texttt{behavior} = f(&\texttt{short\_term\_factors},\\ &\texttt{long\_term\_factors}, \texttt{long\_memory},\\ &\texttt{attacker\_effort}).\end{aligned} \tag{1}$$

Note that f mathematically abstracts and represents the interactions between the four cognitive domains operating in long-memory (i.e., perception, working memory, decision making, and action), while also taking short-term and long-term factors into account. Moreover, f accommodates attacker's effort as input. It is an outstanding research problem to identify the appropriate abstractions for representing these model parameters and what kinds of mathematical functions f are appropriate to what kinds of social engineering attacks. These questions need to be answered using experimental studies. Note that the framework can be expanded to include measures of brain activity, either direct measures such as electroencephalography, or indirectly using peripheral measures such as eye tracking and autonomic nervous system activity (Valecha et al., 2020).

Specific to the cybersecurity domain, we propose considering persuasion-related behavior, as shown in **Figure 2** and the corresponding mathematical representation of Equation (1), meaning that the outcome in Equation (1) can be replaced by the probability that a user is persuaded by the attacker to act as the attacker intended. Intuitively, persuasion is the act of causing someone to change their attitudes, beliefs, or values based on reasoning or argument. Wright et al. (2014) defines Cialdini's Principles of Persuasion, which have been extensively used to study the response to social engineering messages (but not social engineering cyberattacks). **Table 1** presents a brief summary of Cialdini's Principles of Persuasion.

Intuitively, the mathematical function f in Equation (1) should accommodate or reflect Cialdini's Principles of Persuasion. Although the state-of-the-art does not allow us to draw insights into how these Principles would quantitatively affect the form of f, we can still draw some insights from existing studies, as discussed below.

van der Heijden and Allodi (2019) study the relation between phishing attack success and Cialdini Principles of Persuasion using enterprise emails from a financial institution. They find that phishing emails that received the most responses ("clicks") are those who use *consistency* and *scarcity* principles mentioned above. They also find that emails with more cognitive elements (e.g., proper grammar, personalization, and persuasion elements) receive most responses. In a related study, Lin et al. (2019) find that younger individuals are more susceptible to phishing messages that use the *scarcity* and *authority* principles, while

TABLE 1 | Summary of Cialdini Principles of Persuasion.

Principle	Description
Liking	The act of saying yes to something you know and like; for example, a social engineer presenting himself as helpful and empathetic toward the victim in a password reset process.
Reciprocity	Repaying an earlier action in kind; for example, conveying to a victim that they have detected suspicious activities in the victim's credit card account while encouraging the victim to reset the password with their assistance.
Social Proof	The use of endorsement; for example, stating that, due to recent suspicious activities, new security requirements are issued and must be complied by all account holders.
Consistency	Leveraging the desire of individuals to be consistent with their words, belief, and actions; for example, reminding users that they have to comply with a password reset policy as they have previously done.
Authority	Responding to others with more experience, knowledge, or power; for example, an email signed by a Senior Vice President of a bank requesting customers to reset their account passwords.
Scarcity	Something being valuable when it is perceived to be rare or available for a limited time; for example, giving a user 24-h notice before they deactivate the user's account.
Unity	Shared identity between the influencer and the influenced

The principle of Unity was introduced in Cialdini (2016) but has not been studied in social engineering research; it is presented here for the purpose of completeness.

older individuals are more susceptible to phishing emails that use the *reciprocity* and *liking* principles. Rajivan and Gonzalez (2018) find that the most successful phishing message strategies are notifications messages, authoritative messages, friend request messages, shared interest messages, and assistance with a failure. These strategies map to Cialdini's Principles of *liking*, *authority*, and *unity*. Lawson et al. (2018) find that socially engineered messages that use *authority* and *scarcity* principles are considered more suspicious than those that use the *liking* principle.

There have been proposals to augment Cialdini's Principles to better represent the psychological vulnerabilities that have been exploited by social engineering cyberattacks. Ferreira and colleagues (Ferreira and Lenzini, 2015; Ferreira et al., 2015) present five Principles of Persuasion by combining (i) Cialdini Principles of Persuasion; (ii) Stajano's study on scams and how distraction, social compliance, herd, dishonesty, kindness, need and greed, and time affect the persuasive power of scam messages (Stajano and Wilson, 2009); and (iii) Gragg's psychological triggers on how strong affect or emotion, overloading, reciprocation, deceptive relationships, diffusion of responsibility and moral duty, authority, and integrity and consistency can influence an individual's response to social engineered messages (Gragg, 2003). **Table 2** presents a summary of these newly proposed five principles.

Guided by the newly proposed principles, Ferreira and Lenzini (2015) conduct experiments, using phishing emails, to show that *distraction* (e.g., fear of missing out, scarcity, strong affection, overloading, and time) is the most prevalent phishing tactic, followed by *authority* and LSD.

Summarizing the preceding discussion, we draw:

TABLE 2 | Summary of the five newly proposed principles of persuasion that would better represent the psychological vulnerabilities exploited by social engineering cyberattacks (Ferreira et al., 2015).

Principle	Description
Authority	Obeying pretense of authority or performing a favor for an authority.
Social Proof	Mimicking behavior of the majority of people.
Liking, Similarity, and Deception (LSD)	Obeying someone a victim knows/likes, or someone similar to the victim, or someone a victim finds attractive.
Commitment, Reciprocity, and Consistency (CRC)	Making a victim act as committed, assuring consistency between saying and doing, or returning a favor.
Distraction	Focusing on what a victim can gain, need, or lose/miss out.

Insight 13. *The representation of mathematical function f in Equation (1) should adequately reflect the Principles of Persuasion.*

Since quantitatively describing the mathematical function f, as demanded in Insight 13, is beyond the scope of the state-of-the-art, in what follows we explore some qualitative properties of the these mathematical functions, showing how an increase (decrease) in a model parameter would affect the outcome *behavior* (more precisely, *persuasion*) of a victim. These qualitative observations also need to be quantitatively verified by future experimental studies.

3.3.3. Impact of Attacker Effort on Victim Behavior

As shown in Equation (1), the attacker can affect victim behavior through the *attack effort* variable, which can be reflected by attacker's *message quality* and *message appeal* with respect to the victim in question. In terms of the impact of message quality, Downs et al. (2006) find that most individuals rely on superficial elements when determining if a message is legitimate, without knowing that most of those elements can be spoofed. Jansen and Leukfeldt (2016) find that in online banking frauds involving phishing or malware, most victims report that fraudulent stories in phishing emails or phone calls appear to be trustworthy. Message quality appears increase social engineering cyberattack success. Wang et al. (2012) find that *visceral triggers* and *deception indicators* affect phishing responses. Visceral triggers increases phishing responses, whereas deception indicators have the opposite effect, reducing phishing response. Similarly, Vishwanath et al. (2011) find that individuals use superficial message cues to determine their response to phishing messages. The study reports that urgency cues make it less likely for an individual to detect deception cues. One common method of trustworthiness is through the use of visual deception, which is effective because most individuals associate professional appearance and logos with a website or message been legitimate. Visual deception involves the use of high-quality superficial attributes (e.g., legitimate logos, professional design, and name spoofing). Hirsh et al. (2012) find that phishing messages that use visual deception have a higher victim response rate. Jakobsson

(2007) observes that for most individuals, the decision to trust a website or not is based on-site content and not on the status of a site security indicators (e.g., the use security certificates, HTTPS). He also notes that most users could not detect subtle changes in URLs (e.g., a malicious `www.IUCU.com` vs. a benign `www.IUCU.org`). Dhamija (Dhamija et al., 2006) conducts a study on malicious website identification. Using malicious and legitimate websites with professional appearance, with the difference that malicious sites display security indicators (e.g., missing padlock icon on the browser, warning on site's digital certificate), they find that 23% of their participants fail to identify fraudulent sites for 40% of the time. This group of participants are asked to assess a website's legitimacy based on its appearance (e.g., website design and logos), and 90.9% of participants fail to identify a high-quality malicious website that uses visual deception (i.e., URL spoofing replacing letters in a legitimate URL, for example, "W" for "vv").

On the other hand, message appeal is associated with the benefit an individual derives from complying with a request. Halevi et al. (2013, 2015) find that many individuals that fall to social engineering cyberattacks ignore the risk of their actions because they focus on the potential benefit that the phishing email offers. Message appeal has the most weight on social engineering susceptibility. An example of this is the Nigerian scam, also known as the "419" scam. Herley (2012) notes that although the scam is well-known and information on the scam is readily available online, individuals still fall victim to it because the message is designed to appeal the most gullible. Two techniques that are commonly used to increase message appeal are *contextualization*, also known as pretexting, and *personalization*.

- *Contextualization* is a variation of message framing where the sender provides details or discusses topics relevant to the group to vouch for the victim's membership in the group. Luo et al. (2013) conducts a study on phishing victimization with contextualization in the message. In the experiment, they use work benefits and compensation as a pretext in an email to University staff. They find that individuals interpret high-quality content and argument messages (e.g., well written, persuasive messages) as originating from a credible sender. Basing the message argument on a topic that is common within a community (i.e., contextualization) gives the message the appearance of originating from a known person within a group. Using this technique, they are able to achieve 15.24% victimization in 22 min by combining pretexting and message quality. Similarly, Goel et al. (2017) examine the effect of messages contextualization (i.e., pretexting) on phishing message opening and compliance rates. They find that highly contextualized messages that target issues relevant to the victim are more successful. They also find that messages with higher perceived loss have higher success rates than those with high perceived gains. Rajivan and Gonzalez (2018) also find that phishing messages on work-related or social communication topics (e.g., friend requests) have a higher success rate than messages requesting a password change or offering a deal.
- *Personalization* is another framing technique in which a message is tailored to the preference of the victim (Hirsh et al., 2012), such as friendship appeals, expressing similar interests. In a phishing experiment (Jagatic et al., 2007), find that adding personal data found in social networks to phishing emails increased the response rate from 16 to 72%. Rocha Flores et al. (2014) find that targeted, personalized phishing messages are more effective than generic messages. They find that phishing emails that receive the most responses are those perceived to come from a known source. Bullee et al. (2017) also find that emails using personalized greeting line were responded 1.7 times more likely when compared with emails with generic greeting lines.

Summarizing the preceding discuss, we draw:

INSIGHT 14. *Message quality and message appeal, which reflect attacker effort (e.g., using contextualization and personalization), have a significant impact on the attacker's success.*

3.3.4. Countermeasures Against Social Engineering Cyberattacks

There have been some studies on defending against social engineering cyberattacks. First, it is intuitive that effective training should heighten a victim's sense of threat because individuals are more cautious and sensitive to detecting elements that might indicate deception. Along this line, Wright and Marett (2010) find that suspicion of humanity was a dispositional factor that increases the detection of deception in phishing messages, more so than risk beliefs and trust. Pattinson et al. (2012) also find that when individuals participating in experiments are aware that a phishing attack is involved, they perform better on detecting phishing emails. Second, Tembe et al. (2014) find that individuals from the U.S. having higher suspicion and caution attitudes on online communications, when compared to individuals from China and India. Third, Vishwanath et al. (2018) find that habitual patterns of email habits and deficient self-regulation reduce viewers' suspicion. Moreover, detecting deception cues decrease social engineering susceptibility. Kirmani and Zhu (2007) find that detecting persuasion cues in a message activates suspicion and generates a negative response to the message. Fourth, Canfield et al. (2016) find that individuals' inability to detect phishing messages increases their susceptibility, regardless of their cautionary behavior. For suspicions to be effective in reducing social engineering susceptibility, the risk must out weight the benefit of complying with the message (e.g., message appeal) without affecting decision performance (i.e., accuracy, precision and negative prediction) (Cho et al., 2016). Goel et al. (2017) find that suspicion alone does not prevent phishing victimization because they report that individuals that are suspicious about email messages can still fall victim to social engineering cyberattacks. Summarizing the preceding discussion, we draw the following:

INSIGHT 15. *An individual's capabilities in detection of social engineering cyberattacks are affected by their awareness of the* *threats, their cultural backgrounds, and their individual differences in trust/suspicion.*

Insight 15 highlights that there is no silver-bullet solution to countering social engineering cyberattacks. On the contrary, effective defense must take into consideration the differences between individuals because they are susceptible to social engineering cyberattacks at different degrees.

For more effective defenses against social engineering cyberattacks, the following aspects need to be systematically investigated in the future.

- Achieving effective human-machine interactions in defending against social engineering cyberattacks. Effective defense would require to (i) detecting message elements that attempt to increase recipients' trust and (ii) increasing recipients' suspicion on messages. In either approach, we would need human-machine teaming in detecting social-engineering cyberattacks, highlighting the importance of effective defenses against social engineering cyberattacks.
- Improving users' immunity to social engineering cyberattacks. To improve users' immunity to social engineering cyberattacks, we first need to investigate how to quantify their immunity. In order to improve users' immunity to social engineering cyberattacks, we need to enhance users' *protection motivation* and *capabilities in detecting deceptive cues*. One approach is to enhance the user interface design to highlight the security alerts/indicators in email systems and web browsers because their current designs are not effective (Downs et al., 2006; Kumaraguru et al., 2006; Schechter et al., 2007; Abbasi et al., 2016). Along this direction, the user interface must highlight security alerts/indicators with specific and quantified severity of threats to the user. The current user-interface design appears to mainly focus on usability and user experience while assuming the presence of (social engineering) cyberattacks as a default. This design premise needs to be changed to treating the presence of (social engineering) cyberattacks as the default. In order to enhance users' capabilities in detecting deceptive cues, we need to design new techniques to enhance our capabilities in automatically detecting or assisting users to detect social engineering cyberattacks. Automatic detection has been pursued by previous studies, which however only focus on examining certain message elements that are known to be associated with previous social engineering cyberattacks (i.e., signature-based detectors); these signatures can be easily avoided by attackers. In order to possibly detect new social engineering cyberattacks, future detectors should incorporate cognitive psychology elements to detect social engineering cyberattacks, such as the quantification of messages' persuasiveness and deceptiveness. In order to design automated techniques to assist users in detecting social engineering cyberattacks, human-machine interaction is an important issue.
- Achieving human-centric systems design with quantifiable cybersecurity gain. Modern systems design, including security systems design, often focuses on optimizing performance without considering how humans would introduce vulnerabilities while interacting with the system (i.e., assuming away that humans are often the weakest link). One approach to addressing this problem is to change the designers' mindset to treat users of these systems as the weakest link. This can be achieved by, for example, using security designs that are simpler and less error-prone, while considering the worst-case scenario that the users may have a high cognitive load when using these systems. How to quantify users' vulnerability to social engineering attacks is an outstanding problem because it paves the way to quantify the cybersecurity gain of a better design when compared with a worse design.
- Designing effective training to enhance users' self-efficacy. Training is an important mechanism for defending against social engineering cyberattacks. However, it is a challenging task to design effective training. One approach to addressing the problem is to routinely expose users to specific socially engineering cyberattacks (e.g., messages that have been used by attackers). Another approach is to insert sanitized social engineering attacks (e.g., phishing emails without malicious payload) into users' routine activities to trigger users' response to social engineering cyberattacks (e.g., the feedback will point to the user in question whether the user correctly processed the message). This also effectively increases users' perception of social engineering cyberattacks, making them appear more frequent than it is.
- Understanding and quantifying the impact of short-term factors in social engineering cyberattacks. We observe that as discussed above, few studies have examined the effect of short-term factors in social engineering cyberattacks. Short-term factors are known to affect cognition and behavior in other contexts profoundly. As highlighted in Equation (1) and discussed above, we stress the importance of defining and quantifying social engineering cyberattack metrics, which are largely missing and will become an indispensable component of the broader family of cybersecurity metrics as discussed in Pendleton et al. (2016), Cho et al. (2019), and Xu (2019).

Table 3 highlights the five future research directions and their relationships to the insights.

TABLE 3 | Relationships between future research directions and insights.

Future research direction	Insights
Human-Machine interactions	Reduce trust (Insights 1, 2, 10, 14); Increase suspicion (Insights 3, 5, 7, 9)
Immunity to social engineering cyber attacks	Identify and quantify underlying causes of immunity (Insights 5, 7, 9); Security and UI design (Insight 1, 2, 3); Improve message detection (Insight 10, 13, 14)
Human-Centric System Design	Incorporate psychological state of computer user during system design (Insight 1, 2, 3, 11)
Designing effective training	Training based on susceptibility elements (Insight 4, 6, 8, 10, 11, 12, 14, 15)
Understanding and quantifying the impact of short-term factors	Increase research focus on short term factors impact in security (Insight 1, 2, 3)

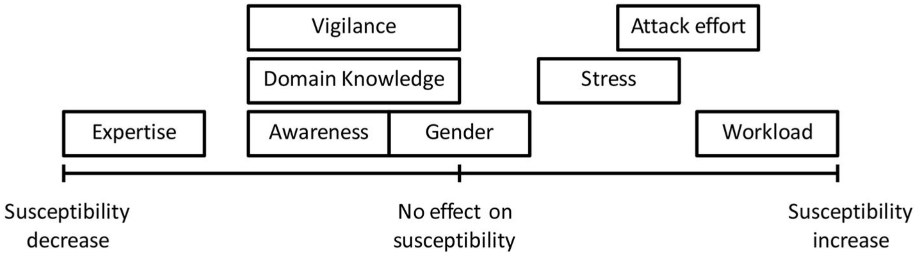


FIGURE 3 | Our *speculation* of the impact on one's susceptibility to social engineering cyberattacks: a factor on the left-hand side means that substantially increasing the factor (e.g., expertise) will decrease one's susceptibility, with further left indicating a bigger degree in decreasing susceptibility; a factor on the right-hand side means that substantially increasing the factor (e.g., workload) will increase one's susceptibility, with further right indicating a bigger degree in increasing susceptibility; gender has little or no effect on one's susceptibility.

3.4. Further Discussion

First, we observe that the 15 insights mentioned above are all *qualitative* rather than *quantitative*. Moreover, the factors are typically investigated standalone. Furthermore, even the qualitative effects are discussed in specific scenarios, meaning that they may not be universally true. Summarizing most of the insights mentioned above, the state-of-the understanding is the following: (i) cognitive workload, stress, and attack effort increase one's vulnerability to social engineering cyberattacks; (ii) the effect of vigilance, personality, awareness, and culture remains to be investigated to be conclusive; (iii) domain knowledge, (certain kinds of) experience, and age (together with certain other factors) reduce one's vulnerability to social engineering cyberattacks; and (iv) gender may not have a significant effect on one's vulnerability to social engineering cyberattacks.

Figure 3 depicts our *speculation* of the impact on one's susceptibility to social engineering cyberattacks. Specifically, we suspect that expertise can decrease one's susceptibility to the largest extent among the factors because expertise equips one the capability to detect the deceptive cues that are used by social engineering cyberattacks. We suspect that vigilance and domain knowledge have a significant, but smaller, impact on reducing one's susceptibility because it is perhaps harder for an expert to fall into victim of social engineering cyberattacks. Since there is no evidence to show which one of these two factors would have a bigger impact than the other, we subjectively treat them as if they have the same impact. We suspect awareness would have a significant impact on reducing one's susceptibility, despite that the literature does not provide any evidence. According to Insight 8 (which is drawn from a body of literature reviewed above), gender has little or no impact. We suspect that stress would decrease one's capability in detecting deception cues, but attack effort would have an even more significant impact on increasing one's susceptibility. We suspect that workload may have the biggest impact on one's susceptibility because it would substantially reduce one's ability in detecting deception cues. For other factors like age and culture, we suspect that their impacts might have to be considered together with other factors, explaining why we do not include them in **Figure 3**. In summary, our understanding of the factors that have impacts on human's susceptibility to social engineering cyberattackers is superficial. This was indeed one of our motivations for proposing the mathematical framework outlined in Equation (1).

Second, it is a fascinating research problem to fulfill the quantitative framework envisioned in the paper because its fulfillment will permit us to identify cost-effective, if not optimal, defense strategies against social engineering cyberattacks. This will also help identify the most important factors. However, we suspect that the optimal defense strategies will vary with, for example, different combinations of short-term factors and long-term factors. For example, we suspect that the importance of factors may be specific to attack scenarios. This is supported by two very recent studies: van der Heijden and Allodi (2019) observed that certain short-term and long-term factors (e.g., workload) may be exploited to wage phishing attacks because malicious emails can coincide with high email volume; and Jalali et al. (2020) showed that certain short-term and long-term factors (e.g., high workload and lack of expertise) are two important factors against medical workers. For example, Insight 6 says that awareness and general technical knowledge do not necessarily reduce one's susceptibility to social engineering cyberattacks; however, this may not hold when taking awareness and human cognition functions into consideration. In other words, we can speculate that the effect of considering one factor alone and the effect of considering multiple interacting factors together may be different. This phenomenon is also manifested by Insight 7, showing that self-efficacy, knowledge, and previous encounter of social engineering cyberattacks collectively reduce one's susceptibility to social engineering cyberattacks. In particular, costly phishing experiences would greatly reduce one's susceptibility to social engineering cyberattacks, while non-costly experiences do not. Putting another way, a certain previous encounter may or may not have a big effect when considered together with other factors.

Third, Insight 12 says that effective training should not ask people to consciously think about social engineering cyberattacks, but making people to formulate an unconscious habit in coping with these attacks. This points out an important research direction on designing future training systems.

Fourth, studies in the context of social engineering cyberattacks inevitably involve human subject, meaning that ethical aspects of these studies must be taken into adequate consideration when designing such experiments and an IRB approval must be sought before conducting any such experiment. For ethical considerations in phishing experiments, we refer to (Finn and Jakobsson, 2007) for a thorough treatment.

4. CONCLUSION

We have presented a framework for systematizing human cognition through the lens of social engineering cyberattacks, which exploit weaknesses in human's cognition functions. The framework is extended from the standard cognitive psychology to accommodate components that emerge from the cybersecurity context. In particular, the framework leads to a representation of a victim's behavior, or more precisely, the degree a victim is persuaded by an attacker to act as the attacker intended, as some mathematical function(s) of many aspects, including victim's cognition functions and attacker's effort. We articulate a number of research directions for future research. We hope that this mathematical representation will guide future research endeavors toward a systematic and quantitative theory of Cybersecurity Cognitive Psychology.

AUTHOR CONTRIBUTIONS

All authors contributed to the article and approved the submitted version.

FUNDING

This paper was supported in part by the Mitre Basic Education Assistance Program (BEAP) and ARO Grant #W911NF-17-1-0566 and NSF Grants #1814825 and #1736209. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The opinions expressed in the paper are those of the authors and do not reflect the funding agencies' policies in any sense.

ACKNOWLEDGMENTS

We thank the reviewers for their constructive comments, which have guided us in improving the paper. In particular, the entire section 3.4 is inspired by the reviewers' comments.

REFERENCES

Abass, I. A. M. (2018). Social engineering threat and defense: a literature survey. *J. Inform. Secur.* 9:257. doi: 10.4236/jis.2018.94018

Abbasi, A., Zahedi, F. M., and Chen, Y. (2016). "Phishing susceptibility: the good, the bad, and the ugly," in *2016 IEEE Conference on Intelligence and Security Informatics (ISI)* (Tucson: IEEE), 169–174. doi: 10.1109/ISI.2016.7745462

Acquisti, A., and Grossklags, J. (2005). Privacy and rationality in individual decision making. *IEEE Secur. Privacy* 3, 26–33. doi: 10.1109/MSP.2005.22

Al'Absi, M., Hugdahl, K., and Lovallo, W. R. (2002). Adrenocortical stress responses and altered working memory performance. *Psychophysiology* 39, 95–99. doi: 10.1111/1469-8986.3910095

Al-Hamar, M., Dawson, R., and Guan, L. (2010). "A culture of trust threatens security and privacy in Qatar," in *2010 10th IEEE International Conference on Computer and Information Technology* (Bradford: IEEE), 991–995. doi: 10.1109/CIT.2010.182

Anderson, R. J. (2008). *Security Engineering: A Guide to Building Dependable Distributed Systems, 2nd Edn*. New York, NY: Wiley Publishing.

Arachchilage, N. A. G., and Love, S. (2014). Security awareness of computer users: a phishing threat avoidance perspective. *Comput. Hum. Behav.* 38, 304–312. doi: 10.1016/j.chb.2014.05.046

Arnett, J. J. (2008). The neglected 95%: why American psychology needs to become less American. *Am. Psychol.* 63:602. doi: 10.1037/0003-066X.63.7.602

Arnsten, A. F. (2009). Stress signalling pathways that impair prefrontal cortex structure and function. *Nat. Rev. Neurosci.* 10:410. doi: 10.1038/nrn2648

Baars, B. J. (1997). In the theatre of consciousness. global workspace theory, a rigorous scientific theory of consciousness. *J. Conscious. Stud.* 4, 292–309. doi: 10.1093/acprof:oso/9780195102659.001.1

Bohm, M. (2011). *Why Russians Don't Smile*. Available online at: https://themoscowtimes.com/articles/why-russians-dont-smile-6672

Bullee, J.-W., Montoya, L., Junger, M., and Hartel, P. (2017). Spear phishing in organisations explained. *Inform. Comput. Secur.* 25, 593–613. doi: 10.1108/ICS-03-2017-0009

Byrne, Z. S., Dvorak, K. J., Peters, J. M., Ray, I., Howe, A., and Sanchez, D. (2016). From the user's perspective: perceptions of risk relative to benefit associated with using the internet. *Comput. Hum. Behav.* 59, 456–468. doi: 10.1016/j.chb.2016.02.024

Cahill, L. (2006). Why sex matters for neuroscience. *Nat. Rev. Neurosci.* 7:477. doi: 10.1038/nrn1909

Cain, A. A., Edwards, M. E., and Still, J. D. (2018). An exploratory study of cyber hygiene behaviors and knowledge. *J. Inform. Secur. Appl.* 42, 36–45. doi: 10.1016/j.jisa.2018.08.002

Canfield, C. I., Fischhoff, B., and Davis, A. (2016). Quantifying phishing susceptibility for detection and behavior decisions. *Hum. Factors* 58, 1158–1172. doi: 10.1177/0018720816665025

Cho, J.-H., Cam, H., and Oltramari, A. (2016). "Effect of personality traits on trust and risk to phishing vulnerability: modeling and analysis," in *2016 IEEE International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support (CogSIMA)* (San Diego, CA: IEEE), 7–13.

Cho, J.-H., Xu, S., Hurley, P. M., Mackay, M., Benjamin, T., and Beaumont, M. (2019). STRAM: Measuring the trustworthiness of computer-based systems. *ACM Comput. Surv* (New York, NY), 51. doi: 10.1145/3277666

Chu, P. C., Spires, E. E., and Sueyoshi, T. (1999). Cross-cultural differences in choice behavior and use of decision aids: a comparison of Japan and the United States. *Organ. Behav. Hum. Decis. Process.* 77, 147–170. doi: 10.1006/obhd.1998.2817

Cialdini, R. (2016). *Pre-suasion: A Revolutionary Way to Influence and Persuade*. New York, NY: Simon and Schuster.

da Veiga, A., and Martins, N. (2017). Defining and identifying dominant information security cultures and subcultures. *Comput. Secur.* 70, 72–94. doi: 10.1016/j.cose.2017.05.002

Damon, W., Lerner, R. M., Kuhn, D., and Siegler, R. S. (2006). *Handbook of Child Psychology, Cognition, Perception, and Language*, Vol. 2. Hoboken, NJ: John Wiley & Sons.

Darwish, A., El Zarka, A., and Aloul, F. (2012). "Towards understanding phishing victims' profile," in *2012 International Conference on Computer Systems and Industrial Informatics* (Sharjah: IEEE), 1–5. doi: 10.1109/ICCSII.2012.6454454

DeValois, R. L., and DeValois, K. K. (1990). *Spatial Vision*, Vol. 14. Oxford University Press. doi: 10.1093/acprof:oso/9780195066579.001.0001

Dhamija, R., Tygar, J. D., and Hearst, M. (2006). "Why phishing works," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY: ACM), 581–590. doi: 10.1145/1124772.1124861

Digman, J. M. (1997). Higher-order factors of the big five. *J. Pers. Soc. Psychol.* 73:1246. doi: 10.1037/0022-3514.73.6.1246

Donnellan, M. B., and Robins, R. W. (2009). "The development of personality across the lifespan," in The Cambridge Handbook of Personality Psychology, eds P. J. Corr and G. Matthews (Cambridge: Cambridge University Press), 191. doi: 10.1017/CBO9780511596544.015

Downs, J. S., Holbrook, M. B., and Cranor, L. F. (2006). "Decision strategies and susceptibility to phishing," in *Proceedings of the Second Symposium on Usable Privacy and Security* (Pittsburgh, PA: ACM), 79–90. doi: 10.1145/1143120.1143131

Elzinga, B. M., and Roelofs, K. (2005). Cortisol-induced impairments of working memory require acute sympathetic activation. *Behav. Neurosci*. 119:98. doi: 10.1037/0735-7044.119.1.98

Evans, J. S. B. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annu. Rev. Psychol*. 59, 255–278. doi: 10.1146/annurev.psych.59.103006.093629

Farhat, N. F. N. (2017). *Scam Alert - Blackmail Email*. Available online at: https://www.linkedin.com/pulse/scam-alert-blackmail-email-ned-farhat

Ferreira, A., Coventry, L., and Lenzini, G. (2015). "Principles of persuasion in social engineering and their use in phishing," in *International Conference on Human Aspects of Information Security, Privacy, and Trust* (Los Angeles, CA: Springer), 36–47. doi: 10.1007/978-3-319-20376-8_4

Ferreira, A., and Lenzini, G. (2015). "An analysis of social engineering principles in effective phishing," in *2015 Workshop on Socio-Technical Aspects in Security and Trust* (Verona: IEEE), 9–16. doi: 10.1109/STAST.2015.10

Finn, P., and Jakobsson, M. (2007). Designing ethical phishing experiments. *IEEE Technol. Soc. Mag*. 26, 46–58. doi: 10.1109/MTAS.2007.335565

Franklin, D. W., and Wolpert, D. M. (2011). Computational mechanisms of sensorimotor control. *Neuron* 72, 425–442. doi: 10.1016/j.neuron.2011.10.006

Gavett, B. E., Zhao, R., John, S. E., Bussell, C. A., Roberts, J. R., and Yue, C. (2017). Phishing suspiciousness in older and younger adults: the role of executive functioning. *PLoS ONE* 12:e0171620. doi: 10.1371/journal.pone.0171620

Gigerenzer, G. (2008). Why heuristics work. *Perspect. Psychol. Sci*. 3, 20–29. doi: 10.1111/j.1745-6916.2008.00058.x

Goel, S., Williams, K., and Dincelli, E. (2017). Got phished? Internet security and human vulnerability. *J. Assoc. Inform. Syst*. 18:2. doi: 10.17705/1jais.00447

Gragg, D. (2003). A multi-level defense against social engineering. *SANS Reading Room* 13, 1–21. doi: 10.1093/acprof:oso/9780199253890.003.0002

Grandstrand, O. (2013). "Cultural differences and their mechanisms," in *The Oxford Handbook of Cognitive Psychology*, ed D. Reisberg (Oxford: Oxford University Press), 970–985.

Gupta, S., Singhal, A., and Kapoor, A. (2016). "A literature survey on social engineering attacks: phishing attack," in *2016 International Conference on Computing, Communication and Automation (ICCCA)* (Greater Noida: IEEE), 537–540. doi: 10.1109/CCAA.2016.7813778

Halevi, T., Lewis, J., and Memon, N. (2013). "A pilot study of cyber security and privacy related behavior and personality traits," in *Proceedings of the 22nd International Conference on World Wide Web* (Singapore: ACM), 737–744. doi: 10.1145/2487788.2488034

Halevi, T., Memon, N., Lewis, J., Kumaraguru, P., Arora, S., Dagar, N., et al. (2016). "Cultural and psychological factors in cyber-security," in *Proceedings of the 18th International Conference on Information Integration and Web-based Applications and Services, iiWAS '16* (New York, NY: ACM), 318–324. doi: 10.1145/3011141.3011165

Halevi, T., Memon, N., and Nov, O. (2015). Spear-phishing in the wild: a real-world study of personality, phishing self-efficacy and vulnerability to spear-phishing attacks. *SSRN Electron. J*. doi: 10.2139/ssrn.2544742.

Harrison, B., Svetieva, E., and Vishwanath, A. (2016). Individual processing of phishing emails: how attention and elaboration protect against phishing. *Online Inform. Rev*. 40, 265–281. doi: 10.1108/OIR-04-2015-0106

Herley, C. (2012). "Why do Nigerian scammers say they are from Nigeria?" in *WEIS* (Berlin).

Hirsh, J., Kang, S., and Bodenhausen, G. (2012). Personalized persuasion: tailoring persuasive appeals to recipients' personality traits. *Psychol. Sci*. 23, 578–581. doi: 10.1177/0956797611436349

Hof, P. R., and Mobbs, C. V. (2001). *Functional Neurobiology of Aging*. Amsterdam: Elsevier.

Hofstede, G. H., Hofstede, G. J., and Minkov, M. (2010). *Cultures and Organizations: Software of the Mind*, 3rd Edn. McGraw-Hill.

Hong, K. W., Kelley, C. M., Tembe, R., Murphy-Hill, E., and Mayhorn, C. B. (2013). "Keeping up with the joneses: assessing phishing susceptibility in an email task," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Los Angeles, CA: SAGE Publications), 1012–1016. doi: 10.1177/1541931213571226

Horn, J. L., and Cattell, R. B. (1967). Age differences in fluid and crystallized intelligence. *Acta Psychol*. 26, 107–129. doi: 10.1016/0001-6918(67)90011-X

Howe, A. E., Ray, I., Roberts, M., Urbanska, M., and Byrne, Z. (2012). "The psychology of security for the home computer user," in *Proceedings of the 2012 IEEE Symposium on Security and Privacy, SP '12* (Washington, DC: IEEE Computer Society), 209–223. doi: 10.1109/SP.2012.23

Hutchins, E. M., Cloppert, M. J., and Amin, R. M. (2011). Intelligence-driven computer network defense informed by analysis of adversary campaigns and intrusion kill chains. *Lead. Issues Inform. Warfare Secur. Res*. 1:80.

Indrajit, R. E. (2017). Social engineering framework: Understanding the deception approach to human element of security. *Int. J. Comput. Sci. Issues* 14, 8–16. doi: 10.20943/01201702.816

Jagatic, T. N., Johnson, N. A., Jakobsson, M., and Menczer, F. (2007). Social phishing. *Commun. ACM* 50, 94–100. doi: 10.1145/1290958.1290968

Jakobsson, M. (2007). The human factor in phishing. *Privacy Secur. Cons. Inform*. 7, 1–19.

Jalali, M. S., Bruckes, M., Westmattelmann, D., and Schewe, G. (2020). Why employees (still) click on phishing links: investigation in hospitals. *J. Med. Internet Res*. 22:e16775. doi: 10.2196/16775

Jansen, J., and Leukfeldt, R. (2016). Phishing and malware attacks on online banking customers in the Netherlands: a qualitative analysis of factors leading to victimization. *Int. J. Cyber Criminol*. 10:79. doi: 10.5281/zenodo.58523

Junger, M., Montoya, L., and Overink, F.-J. (2017). Priming and warnings are not effective to prevent social engineering attacks. *Comput. Hum. Behav*. 66(Suppl. C), 75–87. doi: 10.1016/j.chb.2016.09.012

Kahneman, D. (2011). *Thinking, Fast and Slow*. New York, NY: Farrar, Straus and Giroux.

Kaivanto, K. (2014). The effect of decentralized behavioral decision making on system-level risk. *Risk Anal*. 34, 2121–2142. doi: 10.1111/risa.12219

Kandel, E. R., Schwartz, J. H., and Jessell, T. M. (2000). *Principles of Neural Science*, Vol. 4. New York, NY: McGraw-Hill.

Keil, F. C. (2003). Folkscience: Coarse interpretations of a complex reality. *Trends Cogn. Sci*. 7, 368–373. doi: 10.1016/S1364-6613(03)00158-X

Kenrick, D. T., and Funder, D. C. (1988). Profiting from controversy: lessons from the person-situation debate. *Am. Psychol*. 43:23. doi: 10.1037/0003-066X.43.1.23

Kimball, D. R., and Holyoak, K. J. (2000). "Transfer and expertise," in The Oxford Handbook of Memory, eds E. Tulving, and F. I. M. Craik (New York, NY: Oxford University Press), 109–122.

Kirmani, A., and Zhu, R. (2007). Vigilant against manipulation: the effect of regulatory focus on the use of persuasion knowledge. *J. Market. Res*. 44, 688–701. doi: 10.1509/jmkr.44.4.688

Klein, G. A., and Calderwood, R. (1991). Decision models: some lessons from the field. *IEEE Trans. Syst. Man Cybernet*. 21, 1018–1026. doi: 10.1109/21.120054

Kruger, J., and Dunning, D. (1999). Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *J. Pers. Soc. Psychol*. 77:1121. doi: 10.1037/0022-3514.77.6.1121

Kumaraguru, P., Acquisti, A., and Cranor, L. F. (2006). "Trust modelling for online transactions: a phishing scenario," in *Proceedings of the 2006 International Conference on Privacy, Security and Trust: Bridge the Gap Between PST Technologies and Business Services* (Markham, ON: ACM), 11. doi: 10.1145/1501434.1501448

Lawson, P. A., Crowson, A. D., and Mayhorn, C. B. (2018). "Baiting the hook: exploring the interaction of personality and persuasion tactics in email phishing attacks," in *Congress of the International Ergonomics Association* (Florence: Springer), 401–406. doi: 10.1007/978-3-319-96077-7_42

Lin, T., Capecci, D. E., Ellis, D. M., Rocha, H. A., Dommaraju, S., Oliveira, D. S., et al. (2019). Susceptibility to spear-phishing emails: Effects of internet user demographics and email content. *ACM Trans. Comput. Hum. Interact*. 26:32. doi: 10.1145/3336141

Linvill, D. L., Boatwright, B. C., Grant, W. J., and Warren, P. L. (2019). "The Russians are hacking my brain!" investigating Russia's internet research agency twitter tactics during the 2016 United States presidential campaign. *Comput. Hum. Behav*. 99, 292–300. doi: 10.1016/j.chb.2019.05.027

Luo, X. R., Zhang, W., Burd, S., and Seazzu, A. (2013). Investigating phishing victimization with the heuristic-systematic model: a theoretical framework and an exploration. *Comput. Secur*. 38, 28–38. doi: 10.1016/j.cose.2012.12.003

Lupien, S. J., McEwen, B. S., Gunnar, M. R., and Heim, C. (2009). Effects of stress throughout the lifespan on the brain, behaviour and cognition. *Nat. Rev. Neurosci*. 10:434. doi: 10.1038/nrn2639

Mackworth, N. H. (1948). The breakdown of vigilance during prolonged visual search. *Q. J. Exp. Psychol*. 1, 6–21. doi: 10.1080/17470214808416738

Mather, M., and Sutherland, M. R. (2011). Arousal-biased competition in perception and memory. *Perspect. Psychol. Sci*. 6, 114–133. doi: 10.1177/1745691611400234

McBride, M., Carter, L., and Warkentin, M. (2012). Exploring the role of individual employee characteristics and personality on employee compliance with cybersecurity policies. *RTI Int. Instit. Homeland Secur. Solut*. 5:1.

Mesulam, M.-M. (1998). From sensation to cognition. *Brain* 121, 1013–1052. doi: 10.1093/brain/121.6.1013

Mitnick, K., and Simon, W. L. (2003). *The Art of Deception: Controlling the Human Element of Security*. Indianapolis, IN: Wiley Publishing.

Miyake, A., and Shah, P. (1999). *Models of Working Memory: Mechanisms of Active Maintenance and Executive Control*. Cambridge, UK: Cambridge University Press. doi: 10.1017/CBO9781139174909

Navon, D., and Gopher, D. (1979). On the economy of the human-processing system. *Psychol. Rev*. 86:214. doi: 10.1037/0033-295X.86.3.214

Nisbett, R. E., and Wilson, T. D. (1977). Telling more than we can know: verbal reports on mental processes. *Psychol. Rev*. 84:231. doi: 10.1037/0033-295X.84.3.231

Nosek, B. A., Hawkins, C. B., and Frazier, R. S. (2011). Implicit social cognition: from measures to mechanisms. *Trends Cogn. Sci*. 15, 152–159. doi: 10.1016/j.tics.2011.01.005

Ovelgönne, M., Dumitras, T., Prakash, B. A., Subrahmanian, V. S., and Wang, B. (2017). Understanding the relationship between human behavior and susceptibility to cyber attacks: a data-driven approach. *ACM Trans. Intell. Syst. Technol*. 8:51:1–51:25. doi: 10.1145/2890509

Parasuraman, R., and Rizzo, M. (2008). *Neuroergonomics: The Brain at Work*, Vol. 3. New York, NY: Oxford University Press.

Park, D. C., and Reuter-Lorenz, P. (2009). The adaptive brain: aging and neurocognitive scaffolding. *Annu. Rev. Psychol*. 60, 173–196. doi: 10.1146/annurev.psych.59.103006.093656

Pattinson, M., Jerram, C., Parsons, K., McCormac, A., and Butavicius, M. (2012). Why do some people manage phishing e-mails better than others? *Inform. Manage. Comput. Secur*. 20, 18–28. doi: 10.1108/09685221211219173

Pendleton, M., Garcia-Lebron, R., Cho, J.-H., and Xu, S. (2016). A survey on systems security metrics. *ACM Comput. Surv*. 49, 1–35. doi: 10.1145/3005714

Pfleeger, S. L., and Caputo, D. D. (2012). Leveraging behavioral science to mitigate cyber security risk. *Comput. Secur*. 31, 597–611. doi: 10.1016/j.cose.2011.12.010

Pinker, S. (2009). *How the Mind Works (1997/2009)*. New York, NY: WW Norton & Company.

Purkait, S., Kumar De, S., and Suar, D. (2014). An empirical investigation of the factors that influence internet user's ability to correctly identify a phishing website. *Inform. Manage. Comput. Secur*. 22, 194–234. doi: 10.1108/IMCS-05-2013-0032

Rajivan, P., and Gonzalez, C. (2018). Creative persuasion: a study on adversarial behaviors and strategies in phishing attacks. *Front. Psychol*. 9:135. doi: 10.3389/fpsyg.2018.00135

Redmiles, E. M., Chachra, N., and Waismeyer, B. (2018). "Examining the demand for spam: who clicks?" in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal, ON: ACM), 212. doi: 10.1145/3173574.3173786

Rocha Flores, W., Holm, H., Svensson, G., and Ericsson, G. (2014). Using phishing experiments and scenario-based surveys to understand security behaviours in practice. *Inform. Manage. Comput. Secur*. 22, 393–406. doi: 10.1108/IMCS-11-2013-0083

Salahdine, F., and Kaabouch, N. (2019). Social engineering attacks: a survey. *Future Internet* 11:89. doi: 10.3390/fi11040089

Salthouse, T. (2012). Consequences of age-related cognitive declines. *Annu. Rev. Psychol*. 63, 201–226. doi: 10.1146/annurev-psych-120710-100328

Sample, C., Cowley, J., Hutchinson, S., and Bakdash, J. (2018). "Culture + cyber: exploring the relationship," in *Advances in Human Factors in Cybersecurity, AHFE 2017 International Conference on Human Factors in Cybersecurity*, eds D. Nicholson (Los Angeles, CA: Springer International Publishing), 185–196. doi: 10.4018/978-1-5225-4053-3.ch004

Sawyer, B. D., and Hancock, P. A. (2018). Hacking the human: the prevalence paradox in cybersecurity. *Human Factors* 60, 597–609. doi: 10.1177/0018720818780472

Schaie, K. W. (2005). What can we learn from longitudinal studies of adult development? *Res. Hum. Dev*. 2, 133–158. doi: 10.1207/s15427617rhd0203_4

Schechter, S. E., Dhamija, R., Ozment, A., and Fischer, I. (2007). "The emperor's new security indicators," in *2007 IEEE Symposium on Security and Privacy (SP'07)* (Oakland, CA: IEEE), 51–65. doi: 10.1109/SP.2007.35

Schwabe, L., and Wolf, O. T. (2013). Stress and multiple memory systems: from 'thinking? to 'doing'. *Trends Cogn. Sci*. 17, 60–68. doi: 10.1016/j.tics.2012.12.001

Shaffer, L. (1975). Control processes in typing. *Q. J. Exp. Psychol*. 27, 419–432. doi: 10.1080/14640747508400502

Sharevski, F., Treebridge, P., Jachim, P., Li, A., Babin, A., and Westbrook, J. (2019). Social engineering in a post-phishing era: ambient tactical deception attacks. *arXiv preprint arXiv:1908.11752*. doi: 10.1145/3368860.3368863

Sheng, S., Holbrook, M., Kumaraguru, P., Cranor, L. F., and Downs, J. (2010). "Who falls for phish?: a demographic analysis of phishing susceptibility and effectiveness of interventions," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '10* (New York, NY: ACM), 373–382. doi: 10.1145/1753326.1753383

Shepherd, G. M. (2004). *The Synaptic Organization of the Brain*. New York, NY: Oxford University Press. doi: 10.1093/acprof:oso/9780195159561.001.1

Simons, D. J. (2000). Attentional capture and inattentional blindness. *Trends Cogn. Sci*. 4, 147–155. doi: 10.1016/S1364-6613(00)01455-8

Stajano, F., and Wilson, P. (2009). *Understanding Scam Victims: Seven Principles for Systems Security*. Technical report, University of Cambridge, Computer Laboratory.

Starcke, K., and Brand, M. (2012). Decision making under stress: a selective review. *Neurosci. Biobehav. Rev*. 36, 1228–1248. doi: 10.1016/j.neubiorev.2012.02.003

Tembe, R., Zielinska, O., Liu, Y., Hong, K. W., Murphy-Hill, E., Mayhorn, C., et al. (2014). "Phishing in international waters: exploring cross-national differences in phishing conceptualizations between Chinese, Indian and American sample," in *Proceedings of the 2014 Symposium and Bootcamp on the Science of Security* (Raleigh: ACM), 8. doi: 10.1145/2600176.2600178

Tulving, E., and Craik, F. I. (2000). *The Oxford Handbook of Memory*. New York, NY: Oxford University Press.

Tversky, A., and Kahneman, D. (1974). Judgment under uncertainty: heuristics and biases. *Science* 185, 1124–1131. doi: 10.1126/science.185.4157.1124

Valecha, R., Gonzalez, A., Mock, J., Golob, E. J., and Rao, H. R. (2020). "Investigating phishing susceptibility–an analysis of neural measures," in *Information Systems and Neuroscience* (Vienna: Springer), 111–119. doi: 10.1007/978-3-030-28144-1_12

van der Heijden, A., and Allodi, L. (2019). Cognitive triaging of phishing attacks. *arXiv preprint arXiv:1905.02162*.

Van Schaik, P., Jeske, D., Onibokun, J., Coventry, L., Jansen, J., and Kusev, P. (2017). Risk perceptions of cyber-security and precautionary behaviour. *Comput. Hum. Behav*. 75, 547–559. doi: 10.1016/j.chb.2017.05.038

Vishwanath, A., Harrison, B., and Ng, Y. J. (2018). Suspicion, cognition, and automaticity model of phishing susceptibility. *Commun. Res*. 45, 1146–1166. doi: 10.1177/0093650215627483

Vishwanath, A., Herath, T., Chen, R., Wang, J., and Rao, H. R. (2011). Why do people get phished? Testing individual differences in phishing vulnerability within an integrated, information processing model. *Decis. Support Syst*. 51, 576–586. doi: 10.1016/j.dss.2011.03.002

Voyer, D., Voyer, S., and Bryden, M. P. (1995). Magnitude of sex differences in spatial abilities: a meta-analysis and consideration of critical variables. *Psychol. Bull*. 117:250. doi: 10.1037/0033-2909.117.2.250

Wang, J., Herath, T., Chen, R., Vishwanath, A., and Rao, H. R. (2012). Phishing susceptibility: An investigation into the processing of a targeted spear phishing email. *IEEE Trans. Profess. Commun*. 55, 345–362. doi: 10.1109/TPC.2012.2208392

Wickens, C. D. (2008). Multiple resources and mental workload. *Human Factors* 50, 449–455. doi: 10.1518/001872008X288394

Workman, M. (2008). Wisecrackers: a theory-grounded investigation of phishing and pretext social engineering threats to information security. *J. Am. Soc. Inform. Sci. Technol.* 59, 662–674. doi: 10.1002/asi.20779

Wright, R. T., Jensen, M. L., Thatcher, J. B., Dinger, M., and Marett, K. (2014). Research note-influence techniques in phishing attacks: an examination of vulnerability and resistance. *Inform. Syst. Res.* 25, 385–400. doi: 10.1287/isre.2014.0522

Wright, R. T., and Marett, K. (2010). The influence of experiential and dispositional factors in phishing: an empirical investigation of the deceived. *J. Manage. Inform. Syst.* 27, 273–303. doi: 10.2753/MIS0742-1222270111

Xu, S. (2019). "Cybersecurity dynamics: a foundation for the science of cybersecurity," in *Proactive and Dynamic Network Defense*, Vol. 74, eds Z. Lu and C. Wang (Cham: Springer International Publishing), 1–31. doi: 10.1007/978-3-030-10597-6_1

Zhou, Z., Zhao, J., and Xu, K. (2016). "Can online emotions predict the stock market in China?" in *International Conference on Web Information Systems Engineering* (Shanghai: Springer), 328–342. doi: 10.1007/978-3-319-48740-3_24

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ORIGINAL RESEARCH
published: 02 October 2020
doi: 10.3389/fpsyg.2020.556903

Taking Justice Into Their Own Hands: Predictors of Netilantism Among Cyber Citizens in Hong Kong

Lennon Y. C. Chang[1]* *and Jinxin Zhu*[2]

[1] *School of Social Sciences, Monash University, Clayton, VIC, Australia,* [2] *Assessment Research Centre, The Education University of Hong Kong, Tai Po, Hong Kong*

This research examined the characteristics and predicting indicators of netizens which contribute to "Human Flesh Searching" and internet vigilantism. Human Flesh Searching (HFS) is a form of collective online behavior where netizens contribute information to social media and/or networking platforms about a certain event or a target individual or group to achieve what they regard as justice. It has been used to identify and investigate crime. Some netizens go further and take justice into their own hands by punishing alleged criminals and deviants through online shaming. Using the results of a survey conducted in Hong Kong, the research found both gender and time spent online are not significant variables to predict netizens' intention to contribute to HFS. A positive attitude toward HFS was the strongest predictor of HFS intention. Vigilantism was also a strong predictor of HFS intention. Vigilantism not only affects HFS intention directly, but also indirectly through a positive attitude on HFS. Fairness might negatively influence people's HFS intention and attitude toward HFS; however, this influence was found to be weak in the present study. Social Justice might not affect HFS intention directly, yet it might exert its effect via a positive attitude toward HFS. That is, netizens who intend to contribute to HFS are those who have less confidence in the criminal justice system and believe highly that people should take justice into their own hands.

Keywords: internet vigilantism (netilantism), confidence in criminal justice system, cyber crowdsourcing, social justice, human flesh searching

OPEN ACCESS

Edited by:
Richard Keith Wortley,
University College London,
United Kingdom

Reviewed by:
Peter Nils Grabosky,
Australian National University,
Australia
Kerry John Kennedy,
The Education University
of Hong Kong, Hong Kong

****Correspondence:***
Lennon Y. C. Chang
Lennon.Chang@monash.edu

Specialty section:
This article was submitted to
Cognition,
a section of the journal
Frontiers in Psychology

Received: *29 April 2020*
Accepted: *12 August 2020*
Published: *02 October 2020*

Citation:
Chang LYC and Zhu J (2020)
Taking Justice Into Their Own Hands:
Predictors of Netilantism Among
Cyber Citizens in Hong Kong.
Front. Psychol. 11:556903.
doi: 10.3389/fpsyg.2020.556903

INTRODUCTION

Technology has changed every aspect of our everyday lives. People now do a lot of things through the internet without physical contact. During the COVID-19 lockdown, we saw how people sought to maintain their normal lives without going out. People talked to each other online via social media such as Facebook, Line, WhatsApp, and WeChat. Conferencing apps such as Zoom made it possible for people to organize not only meetings but also parties online. Thanks to these conferencing applications such as Zoom and Cisco WebEx, online teaching and working from home became a "new normal" during the period of lockdown and people even organized virtual social activities such as drinks and parties using new technologies. We also see netizens using the internet, social media and online platforms to investigate crime, to report issues as online journalists and to pass judgment.

Using the skill of cyber-crowdsourcing, "netizens" (citizens actively involved in the online community) can provide information and clues about crime or deviant behavior. Fellow netizens may then conduct further investigations to dig out more information based on the initial information and clues provided. Examples can be seen in online responses which *identify* crime (for example, anti-corruption activities in China), *investigate* crime or deviant behavior (e.g., the 2013 Boston marathon bombing in the United States and police brutality cases in Hong Kong), and/or *punish* criminals through naming and public shaming (e.g., naming and shaming alleged cyberbullies and online child-predators). As Chang and Grabosky (2017: 545) argued, cyber crowdsourcing "has been shown to be a formidable form of private regulation."

Human Flesh Searching (HFS), known as "renrou sousou," or "qi-di" in Chinese, is a good example of how technology is being used to achieve "justice" as perceived by netizens. HFS is a collective online behavior where netizens contribute knowledge and information through social media or networking platforms to expose alleged facts related to certain events and/or to publish information on a target individual or group. It emerged first in China in early 2000 and has become common in the Greater China Region, i.e., the People's Republic of China (China), Hong Kong and Taiwan. Since 2010, it has become common throughout the world (Chang and Poon, 2017). While some HFS is undertaken just for fun or to fulfill one's curiosity (such as gossip about a celebrity), most HFS is undertaken with the aim of exposing crime and deviant behavior, and to shame and punish alleged criminals and deviant individuals (Ong, 2012; Hatton, 2014; Chang and Leung, 2015). Chang and Poon (2017) coined the term "netilantism" (internet vigilantism) to describe the latter behavior.

According to Chang and Poon (2017), netilantism included behaviors such as (1) online activities to identify/disclose crime (such as identifying corrupt officials in China); (2) to investigate crime or deviant behavior (such as netizens trying to disclose the identity of police involved in violent behavior during the 2019 Anti-extradition protests in Hong Kong or 2014 Sunflower Movement in Taiwan); and (3) to punish criminals or deviants through public shaming and naming (such as public shaming of alleged child predators). Social media and networking platforms such as Facebook, Youtube, Weibo, and Telegram are used by internet vigilantes (netilantes) to post information and conduct cyber-crowdsourcing. Traditional police-initiated requests for information from the public (such as America's Most Wanted, Crime Stoppers and *ad hoc* requests) about, for example, the identity of individuals captured on CCTV imagery, do not disclose what information the police have already gathered. The information provided by police is controlled and the information provided to them is not publicly shared. Netilantism differs from this. It provides peer-to-peer, multi-directional information sharing that can be aggregated. We also see that technology and networking platforms are being used increasingly for "sousveillance" in which netizens record and share alleged misbehavior by authorities (Mann, 2004). Although netilantism can contribute to co-production of security and cyber security, it is important to address and mitigate the risks that come with it such as the legitimacy of the information provided, the provision of false or misleading information intended to interfere with or mislead the crime investigation and the consequences that might be caused by identifying the wrong suspect (Chang, 2018; Chang et al., 2018).

Most research on HFS has been focused on HFS in China and has been published in Chinese (Li, 2008; Wang, 2009; Zhu and Liu, 2009). There has also been research on internet vigilantism that categorizes the motives of netilantes (Herold, 2011). Nhan et al. (2017), using the 2013 Boston marathon bombing as a case study, analyzed how cyber-crowdsourcing contributed to the investigation of the event and argued more research needs to be done on the forms and interaction between the police and the public. Recently, a systematic review of HFS cases in the greater China region was conducted by academics in Hong Kong (Chang and Leung, 2015; Chia, 2019a). Chang and Leung (2015) identified differences in types of HFS in Hong Kong, Taiwan, and China. Chia (2019a), using similar methods, reviewed cases in the same region in 2006–2015, through the lens of media studies. Trottier (2017, 2019) argued that weaponized visibility has become a norm in our digital era and proposed a conceptual model of digital vigilantism.

Nonetheless, despite the discussion on the impact of HFS on society and how netizens use HFS to realize their so-called "justice," there are only a few empirical studies examining why netizens contribute to HFS. Skoric et al. (2010), using an online survey with Singaporeans, investigated the relationships between personal characteristics (extroversion, neuroticism, agreeableness, conscientiousness, and openness), Asian values and the contribution to online shaming. Chang and Poon (2017), using empowerment theory, tested the differences between netilantes, bystanders, and victims. Chia (2019b) examined the relationship between media coverage and netilantism and found favorable media coverage is essential to netilantism.

There is still little understanding of why people contribute to netilantism. Do netilantes have similar personal characteristics as vigilantes? Are they engaging in HFS to offset the inadequacy of the formal justice system? Do they have confidence in the current criminal justice system and social justice?

This research will contribute to our knowledge of netilantism from a criminological lens, seeking to understand the relationship between netizens' attitudes toward social justice, fairness and criminal justice systems, and their intention to become netilantes.

The Theory of Planned Behavior (TPB) was developed to predict people's intention to engage in certain behavior. As suggested by the TPB, behavioral intention can be predicted by perceived control, that is, "a person's perception of control over behavioral performance" (Montaño and Kasprzyk, 1997: 71). Montaño and Kasprzyk (1997) indicated that the ease or difficulty of behavioral performance will affect a person's behavioral intention. Guided by the TPB and based on the discussion above, the hypothesized model of this research is presented in **Figure 1**.

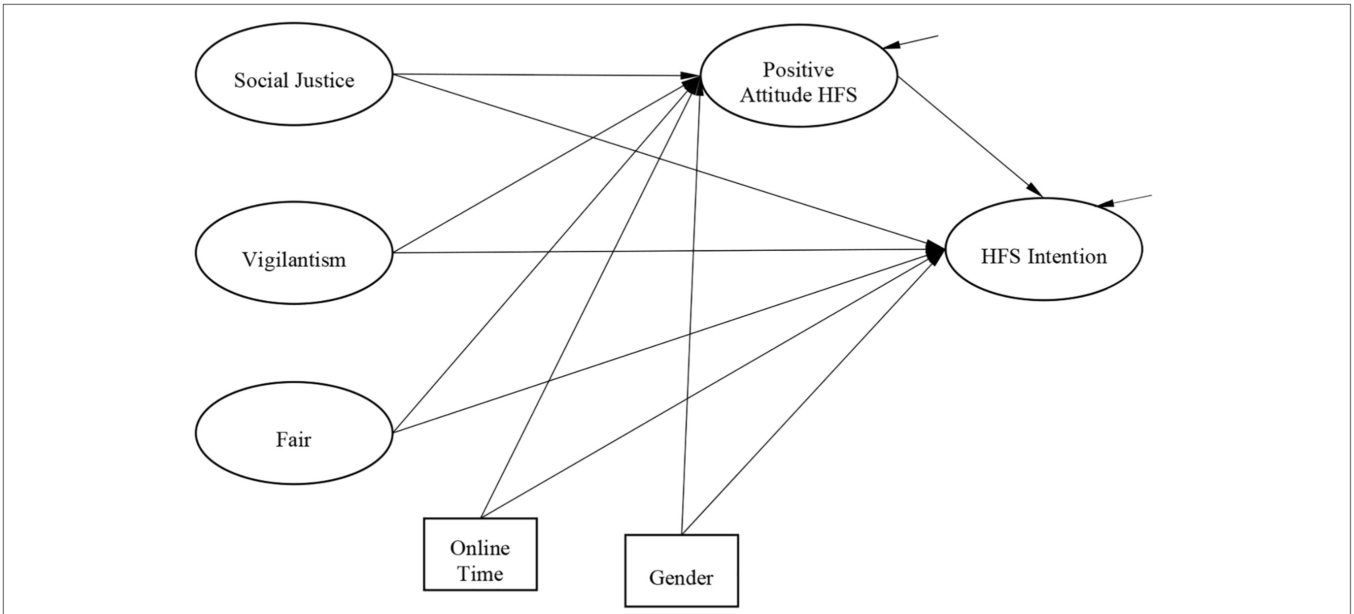


FIGURE 1 | Hypothesized model. Note: Ellipses stand for latent variables and rectangles stand for observed variables.

MATERIALS AND METHODS

Data and Sample

This study used data from a larger study on people's online behavior. The current study focused on HFS. The sample comprised 971 Chinese-speaking respondents in Hong Kong. In the sample, there were 473 (48.7%) male and 492 (50.7%) female respondents. There were 6 respondents (0.6%) who did not provide their gender and they were marked as missing values. The age of the respondents ranged from 14 to 34 years, mainly (93.2%) in the range of 19 to 24 years, and 26 (2.7%) respondents did not provide their age, with the mean age of 21.11 years.

Instruments

The survey questionnaire for the current study was administered in Chinese. It comprised five scales, i.e., Social Justice, Vigilantism, HFS Intention, Positive Attitude toward HFS, and Fair (described in detail below). The scales of Social Justice, Vigilantism, Positive Attitude toward HFS, and Fair all comprised five Likert-type response options, namely, "Strongly Disagree," "Disagree," "Neutral," "Agree," and "Strongly Agree," which were coded as 1,2,3,4, and 5, respectively. The scale of HFS Intention comprised five Likert-type response options of "Strongly Unwilling to," "Unwilling to," "Neutral," "Willing to," and "Strongly Willing to," which were also coded as 1,2,3,4, and 5, respectively. For each question, respondents were instructed to "Please choose one answer and tick where appropriate."

Respondents also indicated their daily online time with nine categories: "Never," "Less Than 1 H," "1–3 H," "3–6 H," "6–9 H," "9–12 H," "12–15 H," "15–20 H," and "More Than 20 H," which were coded as 0 h, 0.5 h, 1.5 h, 4.5 h, 7.5 h, 10.5 h, 13.5 h, 17.5 h, and 22 h, respectively. Both gender and daily time spent online were used as control variables in this research.

Dependent Variable: Human Flesh Searching Intention

Participants were asked their likelihood to contribute to certain HFS activities ("items"). We adopted the items created by Chang and Leung (2015) after reviewing the HFS cases in the greater China region in 2003–2012. The twelve items were:

(1) Corruption activities among government officials;
(2) misconduct of government officials' family members;
(3) sex scandals of government officials;
(4) minor crime issues;
(5) immoral activities;
(6) finding missing people;
(7) helping others to save life;
(8) sex scandals of artists;
(9) incidents about business activities;
(10) expression of personal negative emotions;
(11) helping police to solve certain crimes, and
(12) news about celebrities.

Most of the situations were crime or deviant related scenarios. The Cronbach's Alpha was found in the current study to be 0.934.

Independent Variables

Attitude toward social justice

Five items were used to test participants' attitudes toward social justice. These items were adopted from the Social Justice Scale developed by Torres-Harding et al. (2012) and included:

(1) It is important to make sure that all individuals and groups have a chance to speak and be heard, especially those from traditionally ignored, or marginalized groups;
(2) it is important to talk to others about societal systems of power, privilege, and oppression;
(3) it is important to try to change larger social conditions that cause individual suffering and impede well-being;
(4) it is important to help individuals and groups to pursue their chosen goals in life;
(5) it is important to support community organizations and institutions that help individuals and groups achieve their aims.

The reliability was also tested, and the Cronbach's Alpha was 0.847.

Vigilantism

Participants were asked about their attitude toward vigilantism. Seven questions relating to vigilantism were selected from the confidence of criminal justice systems scales developed by Haas (2010). Participants were asked to answer whether they agree or disagree with statements below:

(1) People who kill armed robbers should not be blamed;
(2) it is sometimes ok for people to take justice into their own hands if they feel the police are unable to protect them;
(3) communities should organize themselves against criminals even if the police disagree with that;
(4) if the government is not successful in their fight against crime, citizens are justified to take the law into their own hands;
(5) citizens should take the law into their own hands more frequently;
(6) it is pointless to hand over a suspected criminal to the police because they will not bring the offender to justice, and
(7) I feel that taking the law into my own hands is justified by circumstances.

The Cronbach's Alpha was 0.860.

Fairness

There were seventeen items used to evaluate participants' attitude toward the fairness of the criminal justice system. Again, they were retrieved from the confidence of criminal justice systems scales developed by Haas (2010). Participants were asked whether they agree to seventeen statements relating to judges and the police:

(1) Judges treat people fairly;
(2) judges are trustworthy;
(3) I can count on the judges to take decisions that are best for society;
(4) I respect judges;
(5) judges deserve respect among citizens;
(6) if a judge passes a light sentence, he will have a good reason for that;
(7) judges' verdicts are well deliberated;
(8) judges do their job well;
(9) judges know what is going on in society;
(10) the police are trustworthy;
(11) the police care about the well-being of every citizen;
(12) I can count on the police to take decisions that are best for society;
(13) the police take citizens seriously;
(14) if the police decide not to arrest someone, they will have a good reason;
(15) the police do their job well;
(16) the police are effective in combating crime, and
(17) the police are there when I need them.

We conducted a two-factor (judges and police) model for the Fairness scale and found that the correlation between these two factors is.55. Also, in the one-factor model, the item loadings were more than 0.5. As a result, the one-factor model was employed in this study. The Cronbach's Alpha was 0.926.

Attitude toward human flesh searching

Six items were used in this research to test participants' positive attitude toward HFS. Chang and Leung (2015) developed the original scale after they reviewed all the literature related to HFS in the Greater China region in 2003–2012. The six items were (1) HFS can maintain justice; (2) HFS can reveal the truth; (3) HFS can punish the bad guys; (4) HFS is very important; (5) HFS can compensate for the inadequacy of the current legal system and, (6) HFS serves justice by neglecting the influence of social hierarchy. The Cronbach's Alpha was 0.850.

Procedure

The data was collected using a face-to-face survey. The survey questionnaire was designed by the research team and was administered in Chinese. The questionnaire interviewers were trained before they started collecting the data. University students in Hong Kong were invited to participate in this survey (see section "Data and Sample"). The survey was conducted one to one or in a small group at university public spaces, mainly at the student canteen. Students participated in this research voluntarily and using their private time. Before the survey started, participants were provided an information sheet describing the project, the interview process, advantages and disadvantages of taking part in the research, information on de-identifying of the data and how the data will be used. The project was approved by the Human Ethical Review Committee at the City University of Hong Kong.

The measurement model was conducted using the multidimensional Graded Response Model (Samejima, 1997) with Mplus (Version 7.2) and the responses to items measuring the five latent variables were specified as ordered categorical. To test the hypothesized model, a two-step analysis was conducted. In the first step, the measurement model was conducted for the five latent variables with daily online time and gender as covariates using Mplus (Version 7.2); meanwhile, 50 sets of plausible values for each latent variable were generated. There were 19 (2.0%) cases with missing values for daily online time or gender. These data were excluded when generating plausible values. The Bayesian estimation approach was adopted for the above mentioned two analyses. In the second step, a path

TABLE 1 | Model constraint information.

Model	AIC	BIC	ABIC	Chi-square (d.f., *P* value)	Chi-square change test (d.f., *P* value)	Estimate with largest *P* value (*P* Value)
Model 1	6783.588	6856.466	6808.827	0.000 (0, 1.000)	N/A	β_{15} = -0.002 (*P* = 0.949)
Model 2	6781.736	6849.756	6805.292	1.609 (1, 0.205)	N/A	β_{25} = 0.004 (*P* = 0.895)
Model 3	6779.977	6843.138	6801.851	0.860 (2, 0.651)	N/A*	B_{24} = -0.013 (*P* = 0.694)
Model 4	6778.385	6836.688	6798.576	0.917 (3, 0.821)	0.057 (1, 0.811)	B_{12} = 0.026 (*P* = 0.442)
Model 5	6777.834	6831.278	6796.342	1.585 (4, 0.812)	0.668 (1, 0.414)	B_{16} = 0.021 (*P* = 0.454)
Model 6	6776.564	6825.150	6793.391	2.145 (5, 0.829)	0.560 (1, 0.454)	β_{23} = -0.081 (*P* = 0.044)

*Chi-square for the baseline model is 290.717 (d.f. = 11, P < 0.001). The estimates were standardized (STDYX). Model 1: Hypothesized model. Model 2: β_{15} = 0. Model 3: β_{15}, β_{25} = 0. Model 4: β_{15}, β_{25}, and β_{24} = 0. Model 5: β_{15}, β_{25}, β_{24}, and β_{12} = 0. Model 6: β_{15}, β_{25}, β_{24}, β_{12}, and β_{16} = 0. NA, Not Available. *The Chi-square value for Model 3 (d.f. = 2) was less than that for Model 2 (d.f. = 1), which suggested that Model 3 was better than Model 2.*

analysis was conducted using these 50 sets of plausible values, as well as the observed values of online time and gender, using Mplus (Version 7.2). By using plausible values, the measurement error was taken into consideration. The standard analysis for plausible value was conducted automatically using Mplus, with the parameter estimates averaged over 50 analyses. However, the indirect effect and total effect of the dependent variables were calculated using the command of "Model Constraint." The following equations describe the hypothesized path model used in the current study:

$$\begin{gathered}\text{HFS Intention } = \beta_{10} + \beta_{11}(\text{Positive HFS Attitude}) + \\ \beta_{12}(\text{Social Justice}) + \beta_{13}(\text{Vigilantism}) + \beta_{14}(\text{Fair}) + \\ \beta_{15}(\text{Gender}) + \beta_{16}(\text{Daily Online Time}) + \varepsilon_1 \end{gathered} \tag{1}$$

$$\begin{gathered}\text{Positive HFS Attitude } = \beta_{20} + \beta_{21}(\text{Social Justice}) + \\ \beta_{22}(\text{Vigilantism}) + \beta_{23}(\text{Fair}) + \beta_{24}(\text{Gender}) + \\ \beta_{25}(\text{Daily Online Time}) + \varepsilon_2 \end{gathered} \tag{2}$$

RESULTS

Effect of Gender and Daily Online Time on HFS Intention and Attitude

A path analysis was conducted to test the hypothesized model using Mplus. The results showed that the hypothesized model was just identified [degree of freedom [d.f.] = 0] and no useful fit information was provided (**Table 1**). To release the degree of freedom, the non-significant effects were fixed at zero with the backward stepwise method based on the largest *P* values. According to the hypothesized Model (Model 1) result, the effect of gender on HFS Intention β_{15} was -0.002, with the largest *P* value of 0.949. Therefore, in Model 2 the β_{15} was fixed at zero. Likewise, β_{25}, with the largest *P* value of 0.895 in Model 2, was fixed at zero in Model 3. By this analogy, all the non-significant coefficients were fixed at zero in Model 6, with β_{23} as the estimate with the largest *P* value of 0.044, which is significant at 0.05 level. The detailed information of the model constraint information is shown in **Table 1**. As is shown in the table, Model 6 was accompanied with the lowest AIC, BIC, and ABIC, which suggested that it was the best model. Also, the Chi-square tests for the change of Chi-square for adjacent models were all non-significant, which indicated later models cannot be rejected. Furthermore, the non-significant Chi-squared test of the model fit for Model 6 showed that the data fitted the model well. This result, on the other hand, showed that gender and daily online time had no significant effect on HFS Intention and HFS Positive Attitude.

Effect of Social Justice, Vigilantism and Fair on HFS Intention and Attitude

The result of the final path model is shown in **Figure 2**, and the total effect, direct, and indirect effect of Social Justice, Vigilantism and Fair on HFS Intention and Positive HFS Attitude are shown in **Table 2**. As is shown in the result, the effect of Positive Attitude toward HFS intention, among the concerned variables, was the strongest (standardized coefficient = 0.500). Vigilantism was also a strong predictor of HFS Intention. The total effect of Vigilantism to HFS Intention is 0.330, with direct effect as 0.154 and indirect effect via Positive HFS Attitude as 0.176. The effect of Fair to HFS Intention was found to be negative (total effect = -0.123, direct effect = -0.083, and indirect effect via Positive HFS Attitude = -0.040). However, no direct effect of Social Justice on HFS Intention was found. Social Justice exerted its effect via the Positive Attitude toward HFS, with a total effect (indirect) of 0.092.

Similarly, Vigilantism was the strongest predictor of Positive Attitude toward HFS (standardized coefficient = 0.354). Social Justice was also a positive predictor, with a standardized coefficient of 0.184. The effect of Fair to Positive HFS Attitude was negative (standardized coefficient = -0.081). In addition, the *R*-squared for Positive Attitude toward HFS was 0.158, and that for HFS Intention was 0.356.

DISCUSSION

From the results, we can argue that netizens who have an intention to contribute to HFS are those who have less confidence in the fairness of the criminal justice system and would take justice into their own hands, irrespective of gender and time spent accessing the internet. Similarly, for those who believe in social justice, if they are provided a tool that they think is efficient for

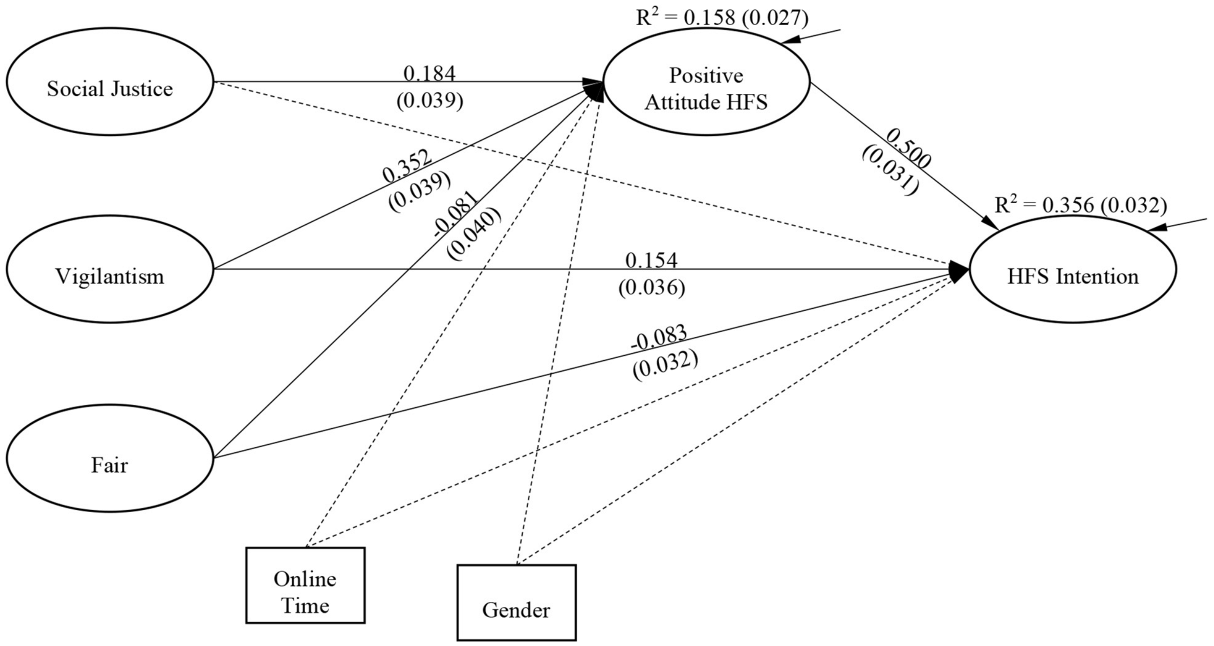


FIGURE 2 | Final model [Fit Indices: Chi-Square = 2.145 (d.f. = 5, p = 0.829), RMSEA < 0.001, CFI > 0.96, TLI > 0.96, and SRMR < 0.01]. Note: 1. All the estimates were standardized (STDYX). The standard errors were presented in parentheses. 2. Two step analysis was conducted. The results of the path analysis shown above were obtained based on 50 sets of plausible value using Mplus. 3. The solid lines indicated significant effects at 0.05 level, and dashed lines indicated non-significant effects at 0.05 level and, therefore, were fixed at zero.

TABLE 2 | Total, direct and indirect effects of dependent variables.

Effect	Estimate	Standard error
Attitude to intention (total effect = direct effect)	0.500	0.031
Social justice to intention (total effect = indirect effect): Indirect effect via attitude	0.092	0.021
Vigilantism to intention (total effect):	0.330	0.036
Direct effect	0.154	0.036
Indirect effect via attitude	0.176	0.022
Fair to intention (total effect):	−0.123	0.036
Direct effect	−0.083	0.032
Indirect effect via attitude	−0.040	0.020
Social justice to attitude (total effect = direct effect)	0.184	0.039
Vigilantism to attitude (total effect = direct effect)	0.352	0.039
Fair to attitude (total effect = direct effect)	−0.081	0.040

All the estimates are standardized (STDYX) and significant at 0.05 level. Attitude = Positive HFS Attitude. Intention = HFS Intention.

them to realize justice, they also will tend to take justice into their own hands.

The results show that those who have less confidence in the criminal justice system are the ones with a higher intention to contribute to HFS and become netilantes. This is aligned with the result of existing research such as Chang and Poon (2017), Chia (2019b), and the concept model developed by Trottier (2019). These are the groups of people who do not have trust in judges and police and believe that people should take justice into their own hands if the legal system cannot protect them. While some of them might already be vigilantes in the real world, the internet provides netizens a new platform to realize the justice which they believe the criminal justice system will not be able to achieve. The intention will be reinforced if they have a positive attitude toward HFS and believe that HFS can help realize justice.

Aligned with the TPB, this research found that a positive attitude toward HFS is the strongest predictor of HFS intention. The HFS platform provides a space for netizens to speak out and contribute to their "justice." Those who believe that the HFS platform provides them with a good way to maintain social justice, reveal truth, punish bad guys, and which can complement the inadequacy of the current legal system have a higher intention to conduct HFS. Indeed, as Gao (2016) argued, the internet has provided a platform for ordinary people to expose information that they were not able to do through traditional media. The HFS platform also provides a good medium for people to pursue their justice outside the traditional criminal justice system, especially for minor local cases that might not receive police attention.

The positive attitude toward HFS also works as a mediator. As mentioned earlier, it empowers those who do not have confidence in their current criminal justice system to take justice into their own hands online. For those who want to build their online reputation, they can publish their identity (real or fake) while disclosing crucial information. As some cases attract attention by traditional media (such as the 2013 Boston marathon bombing case and corruption cases in China), the netilante's contribution to HFS will also be recognized online and possibly also in the media.

The beauty of the HFS platform is that netizens can choose to be identified or to remain anonymous by using a nickname or fake ID. The HFS platform provides those who do not want their real identity to appear on the platform, a channel to provide information. Netizens can hide behind the computer and not have to worry that they will be identified. This might explain why those who tend to have a higher attitude of social justice might not have an intention to contribute to HFS without the mediation of their positive attitude toward the HFS platform. That is, with confidence in the HFS, those who believe in social justice are empowered to contribute without worrying about being identified.

This research shows that for Chinese-speaking respondents in Hong Kong who want to contribute to "justice," technology has provided them a good channel to do so. People, male and female, can take justice into their own hands using the HFS platform. It shows also that not all netizens are netilantes. HFS can be seen as a planned behavior by those netizens who see injustice and unfairness in society and/or who believe they can contribute to realize justice. The HFS platform gives them a good conduit to identify, investigate and even punish a suspect using their own means.

However, it is important that we be wary of the negative effect and ethical concerns that might come with HFS. Cases have already been reported of the wrong person targeted, causing serious damage to the reputation of the person and even leading to suicide (Chang et al., 2018). While HFS can fulfill the public's right to know, it can only be regarded as legitimate when there is a balance between "the public's right to know" and "the individual's right to privacy" (Bu, 2008; Chang and Poon, 2017). As Chang and Poon (2017) argued, "over-justice" of netilantism can develop into a tyranny when the victim's privacy is exploited in an incontrollable manner with no chance for self-defense." As Zetter (2007) argues, activities in cyberspace are too hard to control once they have been initiated. Therefore, while netizens taking "justice" into their own hands might contribute to crime investigation, it is also important to have a second thought before contributing to such activities. There is a need for further studies into mitigation of the damage caused by netilantism. There is also a need for further research to establish whether people who conduct netilantism in western societies have similar characteristics and motivations as those identified in this study of participants in Hong Kong.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Ethics Committee at the City University of Hong Kong. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

Both authors contributed to the article and approved the submitted version.

REFERENCES

Bu, S. T. (2008). A study of renrou sousou and the invasion of privacy. *J. Shenyang Normal Univ.* 32, 93–96.

Chang, L. Y. C. (2018). "Internet vigilantism: Co-production of security and compliance in the digital age," in *Criminal Justice and Regulation Revisited: Essays in Honour of Peter Grabosky*, eds L. Y. C. Chang and R. Brewer (Milton: Taylor & Francis Group), 147–160.

Chang, L. Y. C., and Grabosky, P. (2017). "The governance of cyberspace," in *Regulatory Theory: Foundations and Applications*, ed. P. Drahos (Canberra: ANU Press), 533–551.

Chang, L. Y. C., and Leung, A. (2015). "An introduction to cyber-crowdsourcing (human flesh searching) in the Greater China region," in *Cybercrime Risks and Responses: Eastern and Western Perspectives*, eds R. Smith, R. Cheung, and L. L. Lau (New York, NY: Palgrave), 240–252. doi: 10.1057/9781137474162_16

Chang, L. Y. C., and Poon, R. (2017). Internet vigilantism: attitudes and experiences of university students in Hong Kong. *Int. J. Off. Ther. Comp. Criminol.* 61, 1912–1932. doi: 10.1177/0306624x16639037

Chang, L. Y. C., Zhong, Y., and Grabosky, P. (2018). Citizen co-production of cyber security: self-help, vigilantes, and cybercrime. *Regul. Govern.* 12, 101–114. doi: 10.1111/rego.12125

Chia, S. C. (2019a). Crowd-sourcing justice: tracking a decade's news coverage of cyber vigilantism throughout the Greater China region. *Inform. Commun. Soc.* 22, 2045–2062. doi: 10.1080/1369118x.2018.1476573

Chia, S. C. (2019b). Seeking justice on the web: how news media and social norms drive the practice of cyber vigilantism. *Soc. Sci. Comp. Rev.* 38. doi: 10.1177/0894439319842190 [Epub ahead of print].

Gao, L. (2016). The emergence of the human flesh search engine and political protest in China: exploring the Internet and online collective action. *Med. Cult. Soc.* 38, 349–364. doi: 10.1177/0163443715610493

Haas, N. E. (2010). *Public Support for Vigilantism*. Doctoral thesis, Leiden University, Leiden.

Hatton, C. (2014). *China's Internet Vigilantes and the 'Human Flesh Search Engine*. Available online at: http://www.bbc.com/news/magazine-25913472 (accessed February15, 2020).

Herold, D. K. (2011). "Human flesh search engines: Carnivalesque riots as components of a 'Chinese democracy'," in *Online Society in China: Creating, Celebrating, and Instrumentalising the Online Carnival*, eds D. K. Herold and P. Marlot (Oxford: Routledge), 127–145.

Li, Y. (2008). From violent "human flesh search engine" to friendly "human-computer search engine": a new cognizing the communication value of "human flesh search engine. *Press Circles* 2008, 141–143.

Mann, S. (2004). "Sousveillance: inverse surveillance in multimedia imaging," in *Proceedings of the 12th ACM International Conference on Multimedia*, New York, NY, doi: 10.1145/1027527.1027673

Montaño, D. E., and Kasprzyk, D. (1997). "Theory of Reasoned Action, Theory of Planned Behavior, and the Intergrated Behaviour Model," in *Health behavior and Health Education: Theory, Research, and Practice*, eds K. Glanz, F. M. Lewis, and B. K. Rimer (San Francisco, CA: Jossey Bass), 67–96.

Nhan, J., Huey, L., and Broll, R. (2017). Digilantism: an analysis of crowdsourcing and the Boston marathon bombings. *Br. J. Criminol.* 57, 341–361.

Ong, R. (2012). Online vigilante justice Chinese style and privacy in China. *Inform. Commun. Technol. Law* 21, 127–145. doi: 10.1080/13600834.2012.678653

Samejima, F. (1997). "Graded response model," in *Handbook of Modern Item Response Theory*, eds W. J. van der Linden and R. Hambleton (New York, NY: Springer), 85–100. doi: 10.1007/978-1-4757-2691-6_5

Skoric, M. M., Chua, J. P. E., Liew, M. A., Wong, K. H., and Yeo, P. J. (2010). Online shaming in the Asian context: community empowerment or civic vigilantism? *Surveill. Soc.* 8, 181–199. doi: 10.24908/ss.v8i2.3485

Torres-Harding, S. R., Siers, B., and Olson, B. D. (2012). Development and psychometric evaluation of the Social Justice Scale (SJS). *Am. J. Commun. Psychol.* 50, 77–88. doi: 10.1007/s10464-011-9478-2

Trottier, D. (2017). Digital vigilantism as weaponisation of visibility. *Philos. Technol.* 30, 55–72. doi: 10.1007/s13347-016-0216-4

Trottier, D. (2019). Denunciation and doxing: towards a conceptual model of digital vigilantism. *Global Crime* doi: 10.1080/17440572.2019.1591952 [Epub ahead of print].

Wang, L.-H. (2009). Reflection on "human flesh search. *J. Ning. Rad. TV Univ.* 7, 10–12.

Zetter, K. (2007). *Cyberbullying Suicide Stokes the Internet Fury Machine*. Available online at: http://archive.wired.com/politics/onlinerights/news/2007/11/vigilante_justice (accessed June 26, 2020).

Zhu, B., and Liu, P. (2009). Upset conventional thought: the powerful and horrible "human flesh search engine". *Lib. Inform.* 2009, 102–105.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer, KK, declared a shared affiliation, with no collaboration, with one of the authors, JZ, to the handling editor.

ORIGINAL RESEARCH
published: 15 October 2020
doi: 10.3389/fpsyg.2020.02227

Can You Hear Me Now? Audio and Visual Interactions That Change App Choices

Shakthidhar Reddy Gopavaram, Omkar Bhide and L. Jean Camp*

Luddy School of Informatics, Computing, and Engineering, Indiana University, Bloomington, IN, United States

Android and iOS mobile operating systems use permissions to enable phone owners to manage access to their device's resources. Both systems provide resource access dialogues at first use and per-resource controls. Android continues to offer permission manifests in the Android PlayStore for older apps but is transitioning away from this. Neither manifests nor first-use dialogues enable people to easily compare apps based on resource requests, and the corresponding privacy and security risks. Without the ability to compare resource requests when choosing an app, customers cannot select those apps that request fewer resources. Unnecessary and excessive permission requests, overuse of resources, information exfiltration, and risky apps are endemic. To address this issue we built upon past work in warning science and risk communication to design multimedia indicators to communicate the aggregate privacy and security risk associated with an app. Specifically, we provided participants with a privacy rating using the familiar padlock icon and used audio notifications to either warn or reinforce user choices. We empirically tested participants' app decisions with these padlock icons and audio notifications. The results showed that people with both visual cues and audio feedback are more likely to make app choices that are inversely correlated with the resources requested by the app. Those with neither indicators made decisions reflecting only app rating, while decisions made by those with either the audio or the visual indicators are sometimes inversely correlated with resource requests. This illustrates that simple clear communication about apps' aggregate risk, as opposed to atomic resource requests, changes participants' app selections potentially mitigating the state of information overuse and potential abuse. Additionally, neither the visual indicator nor the audio feedback affected the time required for participants to make a decision.

Keywords: usable privacy and security, human factors, visual indicators, audio indicators, audio warnings, android, permissions manifest, resource access warnings

OPEN ACCESS

Edited by:
*Richard Keith Wortley,
University College London,
United Kingdom*

Reviewed by:
*Ruth Shillair,
Michigan State University,
United States
Mohamed Wiem Mkaouer,
Rochester Institute of Technology,
United States
Mario Linares-Vasquez,
University of Los Andes, Colombia*

****Correspondence:***
*Shakthidhar Reddy Gopavaram
sgopavar@iu.edu*

Specialty section:
*This article was submitted to
Cognition,
a section of the journal
Frontiers in Psychology*

Received: *19 January 2020*
Accepted: *10 August 2020*
Published: *15 October 2020*

Citation:
*Gopavaram SR, Bhide O and
Camp LJ (2020) Can You Hear Me
Now? Audio and Visual Interactions
That Change App Choices.
Front. Psychol. 11:2227.
doi: 10.3389/fpsyg.2020.02227*

1. INTRODUCTION

Apps are often over-privileged, asking for more resources and sharing more information than is necessary. For example, Felt et al. analyzed 940 apps and found that one-third of them were over-privileged, meaning that these apps requested permissions for resources that were beyond what was required for the functionality of the app. Apps requested permissions for system calls they could not use and permissions that had been deprecated (Felt et al., 2011). Such over-permissioning can create a risk to both security and privacy. These risks exist even in apps designed for the most vulnerable users, such as those that are designed for children (Reyes et al., 2017).

Users are responsible for managing risks by approving (or disapproving) app permissions requests in both iOS and Android devices. That users are responsible for making these decisions does not mean that they have the ability or incentives to make informed decisions that accurately reflect their preferences. Informed decision-making requires that users understand permissions and their implications. Yet past research has shown that users do not comprehend the permissions much less their implications (Felt et al., 2012; Kelley et al., 2012; Agarwal and Hall, 2013). Additionally, Some risks cannot be determined by resource requests alone; for example, determining which photo app implements editing on the cloud (along with the security of the remote copies) requires focused technical research (Pan et al., 2018). To evaluate apps nontechnical people are relying on peer patterns of use, social feedback, ratings, and Android market reviews. These do not include usable information about over-privileging, use of resources, or corresponding risks.

One approach to mitigate information exfiltration risk is to implement a machine learning model that predicts user preferences and takes appropriate action at runtime (Olejnik et al., 2017). While a machine learning approach can reduce risk by obfuscating or denying access to sensitive resources, it does have some drawbacks. For one, this approach does not address how an app uses the information it collects from the user. For example, once a user provides an app with certain information he/she may not be able to prevent the app from sharing that information with third parties. Additionally, obfuscating techniques may not be effective at protecting user privacy (Shokri et al., 2010), and denying access to certain permissions can render the app unusable. Therefore, a method for communicating risk at the point of sale is still needed to support risk-aware decision-making (Patil et al., 2016). Specifically, it is important to communicate the aggregate privacy risk arising from different sources like permission requests and data usage practices and communicate it to the user at the time of app selection. Such communication of risk at the time of app selection would help participants select privacy-preserving applications while avoiding the above-mentioned issues.

In this paper, we build upon past research in risk communication to design indicators that communicate the aggregate privacy risk to the user at the time of app selection. We provide cognitively simple visual indicators to communicate the aggregate risk associated with an app to address the problem of information asymmetry and user comprehension. We added negative audio feedback to alert users about potentially high risk apps. Similarly, we implemented positive audio feedback for selecting low-risk apps. This audio feedback in combination with visual cues resulted in participants making app choices that are a function of the indicated risk level. We grounded our experiment in previous research on decision-making in psychology as well as in research in warnings and indicators from offline risk communication.

The innovation in this paper is the combination of aural cues and visual icons that prove efficacious in terms of changing decision-making. The goal of this work is to empower users to choose apps based on the implicit risk that is embedded within the app design and resource requests. The underlying assumption is that it is feasible to estimate the risk of an app given the state of art in mobile security and the requirement for apps to explicitly state their resources requests. We provided aural feedback in the form of cheers and jeers in addition to a standard visual icon for security. Not only could participants easily comprehend the positive nature of joyous cheers and the negative implication of angry jeering without any additional cognitive effort, but they are also not interrupted in the app selection task (no additional clicks or screens are needed). Our results showed that participants with both visual and aural cues were more likely to make app choices corresponding to lower risk exposure. The icons, sound files, and JavaScript that implemented the experimental store as well as details on our Institutional Review Board approval are available upon request.

In the immediately following section, we ground our experiment in the existing permissions models, their drawbacks, and the different factors that affect an individual's comprehension of permissions, potential risks, and corresponding decision-making process. Sections 3, 4 give a detailed description of the experiment. Section 5 provides the results and analysis, followed by a discussion of the possible implications of our findings. We close with our conclusion and possible future work with a focus on the interdisciplinary.

2. BACKGROUND

Here we ground our experiment in the user understanding of the permissions models and corresponding potential risks at the time of the work for Android and iOS. We also discuss the implications for the choice of both systems. For both platforms, the two operating systems automatically grant apps permissions to resources that pose very little risk while requiring explicit human interaction to access more sensitive resources. Android has traditionally provided install time permissions manifests. The decision-maker had the option to install the app and grant it all the permissions in the manifest, or they could deny the permissions and not install the app. This is still the case for devices running Android 5.1 or lower. For Android 6.0 and higher, Google is moving toward the more granular run-time iOS model. In the iOS model (and Android versions 6.0 and higher), people are presented with permissions requests during run-time. The first time an app attempts to access a resource (e.g., location), the system generates a resource access warning. These resource access warnings are similar to warning dialogs on other platforms. People also have the option to revoke permissions that were previously granted by navigating to Privacy Settings in iOS or Application Manager in Android. While iOS's model enables setting custom permissions for each app, research has indicated that it fails to provide users the flexibility they desire (Benisch et al., 2011). Prior research has also found that the iOS vetting and run-time warnings were less effective than Android's community ratings and permissions manifest mechanism (Han et al., 2014). A side-by-side comparison of 2,600 apps offered by the same third parties on the two different platforms (e.g., Uber Android vs. Uber iOS) found that the iOS versions consistently access more resources and exfiltrated more data when compared to their Android counterparts (Han et al., 2013). Therefore, expecting the replacement of the Android permissions model with the iOS model to address users' privacy challenges seems unduly optimistic.

2.1. Drawbacks of Existing Permissions Models

Neither of the two permissions models has proven to be successful in providing consumers with actionable information for making informed decisions (Agarwal and Hall, 2013). Therefore, both iOS and Android users are largely unaware of the resources accessed by the apps (Mylonas et al., 2013). One of the reasons for this is the users' habituation to ignore the current interactions presented in both Android and iOS permissions models. In the case of textual warnings or permissions manifests used in Android, past research has shown that people usually ignore or pay little attention to them (Felt et al., 2012). More specifically, a series of online surveys and laboratory studies conducted by Felt et al. found that only 17% of the participants paid attention to permissions during app installation (Felt et al., 2012). Consumers are also accustomed to ignoring resource access warnings. Warning dialogs are excessively used in today's computers and mobile devices. This overuse of warning dialogs has desensitized people toward them. Therefore, people view these warning dialogs as interruptions rather than security/privacy alerts and click through them to get on with their current task (Xia and Brustoloni, 2005; Brustoloni and Villamarín-Salomón, 2007; Egelman et al., 2008; Sunshine et al., 2009).

Users' inability to comprehend the permissions presented to them and their implications is another reason why the current permissions models are unsuccessful. Textual warning in permissions manifests, for example, are commonly requested in English with too much technical jargon which effectively assumes that all smartphone users possess an above-average level of basic literacy in addition to computer literacy required to comprehend the permissions information and translate to the risks of agreeing to the requested permissions. However, this is not the case. Not all smartphone users have basic education or computer literacy. As a result, they do not understand the technical jargon used to describe permissions or the implications of providing sensitive permissions to applications (Felt et al., 2012; Kelley et al., 2012). Therefore, even though people value their online privacy (Nissenbaum, 1998), they are unable to make privacy-preserving decisions as the current permissions models fail to provide them with actionable risk information.

In recognition that the previous permissions models were inadequate, there has been a move to automate permissions decisions based on observed user behavior. Models of user preferences may be driven by background observations, possibly augmented by explicit queries about acceptable data use (Olejnik et al., 2017; Wijesekera et al., 2017). Such controls can limit resource use by apps but do not enable apps to compete in the marketplace for risk-averse users. Machine learning mitigates risk, but even those people who value their privacy are unable to make privacy-preserving app selections as there is not adequate decision-making support when needed (Papacharissi and Zizi, 2010). Later automated support to constrain resource use is valuable. Yet, a privacy-seeking user may, for example, accidentally choose a photo or audio app which cannot function without the content being sent to the cloud over a more desirable app unless the information is provided in an easy to comprehend manner at the moment of app selection.

2.2. Privacy Indicators

As mentioned above, not everyone has the basic education and computer literacy to understand the information presented in the privacy warning and the risks of giving access to sensitive resources. In such cases, simple privacy indicators that summarize the privacy risks can be beneficial. Locks have been found to have the greatest impact on decision-making in the mobile context (Rajivan and Camp, 2016; Momenzadeh et al., 2020) and communicating security on the web (even when that communication is incorrect; Kelley et al., 2018). Another option for risk indicators, particularly for privacy risk, is the use of eyes as a social cue for information exposure. This has had mixed results. Schlegel et al. (2011) used eyes on the home screen of a smartphone to represent the number of accesses granted to a user's location. The size of the eyes corresponded to the number of times the location was accessed. Liccardi et al. (2014) used eyes to communicate sensitivity score (like our five lock score here) and highlight risky permissions in Android's permissions manifest. Liccardi et al. found that the implicit ranking combined with eyes resulted in significant statistical changes, but he did not compare this with other modes of communication.

Eyes have not consistently proven to be effective or to communicate risk. For example, Benton et al. (2013) compared text with eyes to determine their relative efficacy in communicating aggregate privacy risk to users. Their findings show that eye icons had a stronger statistically significant result when compared both with standard text warnings and brief simplified textual warnings. Yet, using the same eye icons as the previous work, the researchers found that there was no consistent relationship between the impact of the eye icon's effect and the selection of more or less risky apps when roughly accurate ratings were provided using eyes at decision time (Benton et al., 2013).

In a direct comparison between different types of privacy icons in a mobile marketplace, Rajivan et al. studied the effectiveness of three different visual indicators (frowning faces, eye icons, and lock icons), and different framing (positive and negative framing) to evaluate their effect on changes in app selection. The eye icon and face icons were presented with negative framing, as with Liccardi (Liccardi et al., 2014) and Schlegel (Schlegel et al., 2011). The locks were presented as a gain, aka positive framing. The results of the comparison across three icons showed that participants who were presented with positive framing using the padlock made app choices that consistently aligned with increased privacy (Rajivan and Camp, 2016). The impact of the lock icon was significant across all app categories as opposed to the eye icon or the faces. The confidence significantly increased in the presence of priming. Therefore, in our work, we use the lock icons and sought to provide priming with the addition of audio feedback.

2.3. Framing of Privacy

Researchers also explored positive and negative framing and how it affected user decisions. Positive framing refers to communicating security as a benefit that is gained rather than security as something that enables loss avoidance. Positive framing is generally supported by work in the psychology of security, although it has been less often applied in the case of mobile marketplaces (Acquisti et al., 2015). West in 2008 identified the underlying human decision-making biases which imply that gain framing would be more effective than loss framing in communicating computing behaviors (West, 2008). Garg expanded on the previous work, focusing on examples comparing loss versus gain framing specifically in computer security (Garg and Camp, 2013). Anderson and Moore (2009) also noted the power of positive framing security information.

In contrast, Choe et al. (2013) initially found limited efficacy for either framing, with little difference between positive and negative framing in an initial study. In a later study, the same authors reified the consensus that the framing of visual cues could affect participants' permissions-based app decisions. That effect was measured by presenting participants with the same app repeatedly and by asking them to make a comparison between two scales (one negative and one positive). The study found that participants made more risk-averse choices with positive framing (Chen et al., 2015).

2.4. Timing

Timing also influences user attention to warnings. Balebako et al. investigated the ability of users to recall permissions notices when they were presented under three conditions in the app store: when an app was launched, during app use, and after app use. They used recall as a measure of user attention. Their results showed that people paid more attention to permissions when they were presented during app use (Balebako et al., 2015). Their results also showed that users are unlikely to pay attention to permissions shown in the app store. A difference between that work and ours is that informed decision making, not recall, is the focus of our work.

In contrast, Kelley et al. (2013) found that when permissions were included in the app description page instead of being presented after people chose to install an app, people chose apps that had fewer permissions. In that study, they asked participants to imagine that they were choosing the apps for a friend. We know from risk science that people are more accurate in their risk estimates when making judgments about the acceptability of risk for others. In general, people have been found to be more impartial and risk-averse while recommending a risky situation to others (Helfinstein et al., 2015). Availability, affect, assimilation and representativeness can all result in different estimates for privacy risk for oneself when compared to a friend (Garg and Camp, 2013). Thus, the more risk averse behavior may stem from the experiment design as well as the presentation of permissions. In our study, we used app selections for self, and we minimized the cognitive requirements for our participants by using icons and sound.

2.5. Generating Privacy Ratings

Although the generation of accurate Privacy Ratings is not the focus of our research, the possibility of doing so underlies the entire experiment. Therefore, here we provide a shortlist of related work to show that generating such ratings consistently is possible; but not to argue for any of these. Researchers at Carnegie Mellon University have created a website privacygrade.org which gives Android apps a Privacy Grade based on both static code analysis and crowd-sourcing (Lin et al., 2012, 2014). Static code analysis determines what permissions are accessed by an app while the crowd-sourcing aspect determines if the permissions accesses meet user expectations. For example, it is reasonable for Google Hangouts to access a microphone but it would be odd for Angry Birds to do so. It is also possible to rate privacy by analyzing privacy policies. This was demonstrated for websites by Privacy Finder and Privacy Bird (Byers et al., 2004; Cranor et al., 2006; Mcdonald et al., 2009; Tsai et al., 2011). Another promising avenue is the use of natural language processing (NLP) to analyze app description (Pandita et al., 2013). Others have proposed a combination of permission-based risk signals and machine learning techniques to generate a privacy rating (Gates et al., 2014). More thorough evaluations of data flow (e.g., Egele et al., 2011; Pan et al., 2018) and detailed analyses could also be used to develop consistent app ratings (e.g., Beresford et al., 2011; Enck et al., 2011, 2014; Zhou et al., 2011; Arzt et al., 2014).

3. METHODS AND DESIGN

The goal of our work is to see if providing aggregate risk information in form of visual cues (using padlock icons), aural communication, or an integrated warning system containing both would result in users changing their selection of mobile apps. We describe the icons and the sound in detail in this section, grounding them in the previous work from above.

We align our design with the five principles proposed by Rajivan and Camp (2016). Here, we quote directly his conclusions about risk communication. First, "icons should be presented early in the decision-making process while people compare apps to choose and install." Second, "the scale of privacy communicating icons should be consistent with other indicators." In this case, the other indicators are rating and download counts. Third, "privacy communicating icons should be in terms of privacy offered by the app/software." We are evaluating icons for risk, which include privacy and security. Thus we selected a widely used risk communication icon. That we did this is in part based on Rajivan's fourth principle, "icons should align with user mental models of security." Finally, his fifth recommendation is on requirements for the validity of the underlying rating. This does not apply for this experiment as the risk values are randomly assigned during the experiment to mitigate familiarity issues and more subtle biases from, for example, more attractive app icons.

Much previous work has found that priming for privacy has a significant impact on privacy behaviors, but this priming is not feasible in daily practice (Acquisti et al., 2015). To return to the previous example, Rajivan and Camp (2016) illustrated that the greatest effect in app selection occurred when there was both the lock icon and priming for privacy. Grounded in these findings we used two kinds of interactions: one enables comparisons during app selection and the other functions as a warning or validation before installation. The first is a commonly used visual indicator for security and privacy. It provides a simple and easy way to communicate a summary of risk (e.g., resource requests) across apps in one category. The second, a sound notification as a warning, is also designed to serve to prime users for privacy. Building on the study of hazards and warnings, the icon is intended to provide information processing support while the audio is more aligned with warnings as transmission or alert (Wogalter et al., 2005). The combination of these two messages is designed to create a warning system that addresses both the consumer's right to know (with visual decision support) and the duty to warn (with audio installation warnings) that are at the core of risk communication (Viscusi and Zeckhauser, 1996).

We designed the experiment to measure the effectiveness of the two interactions individually and the combination of them in a warning system. Testing this integrated warning system also requires evaluating each individual component. The control enabled us to compare the discrete components and the entire warning system with previous approaches. In this section, we provide detailed information about the two interactions, the four groups of subjects, and the controlled environment.

3.1. Visual Indicator

The goal of the visual indicator is to provide users with easy-to-understand privacy information. A simple icon can ideally inform people with varying levels of literacy. Building upon previous research in this area (discussed in section 2), we employed positive framing using the padlock icon. The design also embeds the standard rubric that when there is a highly variable audience, warnings should be designed for the low-end extreme to include the entire population (Wogalter et al., 2005).

Based on the goal of providing positive framing, more locks imply that an app is associated with lower risk, something that is traditionally indicated through resource requests. In **Figures 2A**, **3A**, we show the lock icon in the context of the *list of apps* page and the *app description* page.

3.2. Audio Feedback

The visual icons provide decision support when users are processing information about the apps. The sound provides feedback (a warning or a verification) to the user immediately after selecting an app. The use of sound notifications is both a practical approach to priming and is consistent with the use of tones for creating immediate human responses to potential hazards (Mileti and Sorensen, 1990).

The use of audio in this experiment builds on both warnings research and past human-subjects research in privacy, specifically research involving priming. Users generally make more privacy-preserving decisions when they are primed for privacy, as noted in section 2. However, a common approach to prime for privacy is to use a survey. Questionnaires for app installations in the real world are not workable. Thus we embedded priming in the experiment as an alert consisting of audio snippets of cheers or jeers. The cheers are played when a person selects an app with a high Privacy Rating (privacy-preserving app) and the jeers are played when a person selects an app with low privacy rating (privacy-invasive app). The cheers were intended to encourage people to select more privacy-preserving apps. The jeers, on the other hand, were intended to warn people about privacy-invasive apps.

We played the audio feedback when a participant selected an app from the *list of apps* page and was transitioning to the *app description* page. An illustration of this is shown in **Figure 1**. Therefore, these notifications do not create any additional tasks or interrupt the app installation process.

3.3. Experimental Groups

To measure how the visual indicators and the audio feedback change users' behavior, we conducted a between-subjects experiment with four experimental groups. There was one control group and three experimental groups: Lock Group, Sound Group, and Warning System Group. The participants in all four groups were presented with a PlayStore simulator which was modeled after Google's PlayStore and simulated the interactions required to install apps on an Android device. However, participants in the experimental groups had additional features available to them. People in the Lock Group were provided with visual indicators for aggregate privacy rating. The participants in the Sound Group heard sound notifications but did not have visual indicators. Finally, the participants in the Warning System Group were provided with visual indicators and were primed for privacy using sound notifications. **Table 1** provides the list of features available to each group.

3.4. Experimental Platform

The experimental platform was an interactive PlayStore simulator. Since we are testing aural feedback and decision support to understand the change in behavior caused by the proposed interactions, it is important for us to trigger the decision processes involved in real-world app installations. In To do so, we built an interactive PlayStore simulator modeling Google's PlayStore. The simulator ran on a web browser and provided identical controls and navigation.

The simulator consisted of three critical components: the *list of apps* page, the *app description* page, and the flow between them. The *list of apps* page models the interface used by the PlayStore to display apps by category. For this experiment, we produced two versions of the *list of apps* page. One version, shown in **Figure 2B**, provides users with just the App Rating. This version is used for the Control and Sound groups as participants in these groups are not presented with visual indicators. The alternative version, shown in **Figure 2A**, augments the *list of apps* page with visual indicators for Privacy Rating in addition to the App Rating. This version is used for experimental groups that provide users with Privacy Rating, i.e., Lock and Warning System groups. In both versions, we only display eight apps per category and when a user selects an app by clicking on it, he/she is redirected to the app description page.

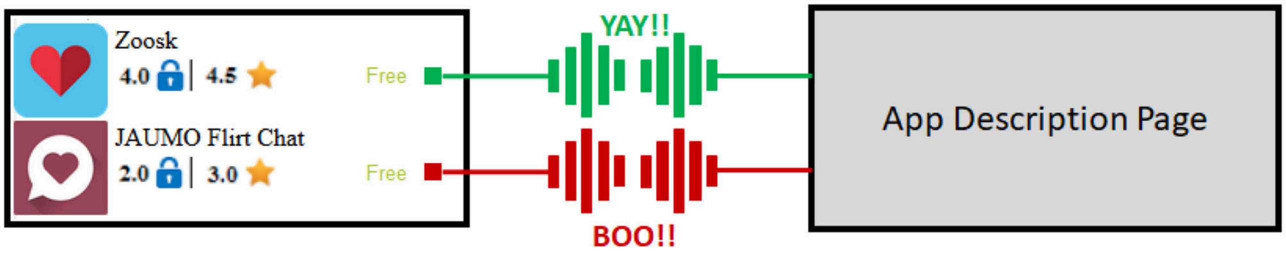


FIGURE 1 | Sounds associated with apps. Choosing an app with a high privacy rating results in cheers and choosing an app with a low privacy rating results in jeers.

TABLE 1 | List of features available in different experimental groups.

Privacy cues	Group 1: control	Group 2: lock	Group 3: sound	Group 4: warning system
Permissions manifest	Yes	Yes	Yes	Yes
Padlock privacy rating	No	Yes	No	Yes
Audio feedback	No	No	Yes	Yes

The app description page on the PlayStore provides users with app rating, download count, a permissions manifest, and an install button. Similar to the *list of apps* page, the *app description* page has two versions: one version with visual indicators for privacy (**Figure 3A**) and the other without (**Figure 3B**). The *app description* page without visual indicators for privacy was shown to participants in the Control and Sound groups. The *app description* with privacy visual indicators was shown to participants in the Lock and Warning System groups. For all four experimental groups, clicking on the install button would mimic the installation of the application.

Additionally, for the Sound and Warning System groups, the simulator plays sound notifications after app selection. These sound notifications are played when a user selects an app in the *list of apps* page and is transitioning to the *app description* page. An illustration of this is shown in **Figure 1**.

All participants were able to navigate the simulator as if in the PlayStore. Specifically, participants were able to move back and forth between the above-mentioned pages using the back arrow, as well as install apps, uninstall apps, and view the permissions manifest by clicking on the *click to view all permissions* dropdown.

3.5. Apps

In this experiment, we selected dating and puzzle apps that were popular at the time of the experiment. We derived a total of 16 apps (8 apps per category) from the PlayStore using the top charts filter for each category.

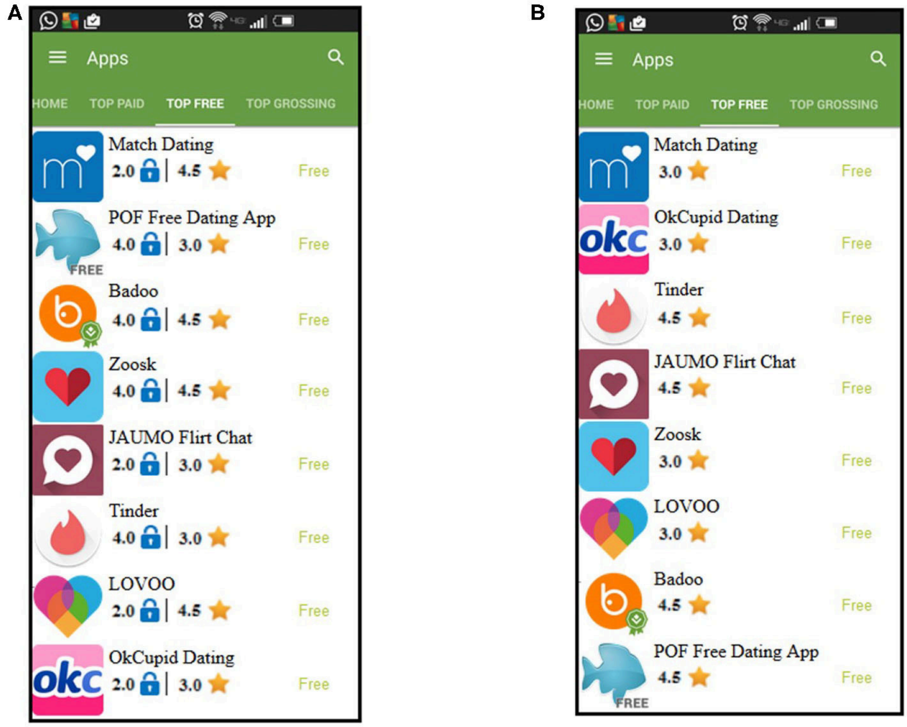
One decision about app selection that varies from previous research is the method of addressing familiarity. Familiarity and reputation are consistently factors in trust decisions in a wide range of online environments (Costante et al., 2015). A series of surveys, interviews, and focus groups illustrated that nontechnical users consistently believe that popularity indicates the acceptability of privacy policies with use by others being an implicit, environmental cue (Morton, 2014). Familiar technologies were found to be perceived as less risky in an investigation of risk perception in mobile and wearable devices (Lee et al., 2015). Specifically, in the case of smartphone applications, past research has shown that users rely on familiarity and majority vote (App Rating) to make app choices (Joeckel et al., 2017). That being the case, it is critical that any interventions introduced to encourage users to make privacy-preserving app choices should be effective in the presence of popular/familiar applications.

Choosing the inclusion of familiar apps required that the experiment design address the potential bias created by familiarity and reputation. In order to mitigate the biases from familiarity and reputation, we randomized the assignment of values for experimental variables for each and every participant, i.e., the values attributed to the apps will vary from participant to participant. As shown in **Figure 4**, the Privacy Rating for the OkCupid Dating app is different for participants 1 and 2. **Figure 4** shows that seven out of the eight applications have different sets of values for Privacy Rating and App Rating. Therefore, if people keep selecting similar applications because they are familiar with them, then there will not be statistically significant differences between the control group and the experimental groups. We would only find the data to be statistically different if people in the control group make decisions based on different experimental variables when compared to the people in the experimental groups. The difficulty in controlling for familiarity was one reason we choose to recruit a large number of subjects in each category.

We also randomized the order of apps and categories to remove any bias caused by ordering.

3.6. Experimental Variables

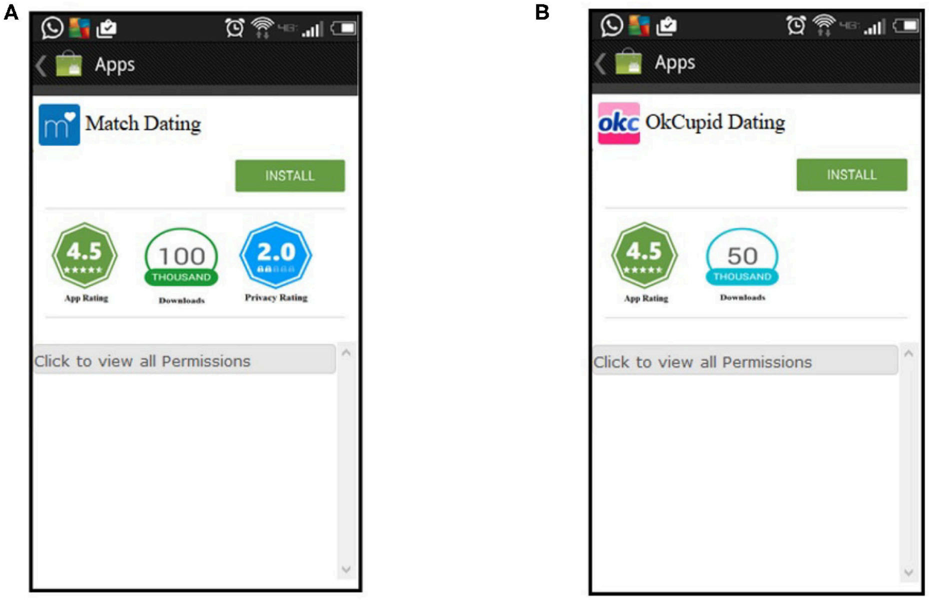
For each app installed by a participant, we recorded the values for Privacy Rating, App Rating, and Download Count. By recording these values we were able to measure the influence they had on the participants' app choices at the time of app selection. In addition to the three experimental variables, we also compute two other variables PrivacyOverAppRating and PrivacyOverDownloadCount. These two additional experimental variables measure the difference between Privacy Rating and the



Match Dating 2.0 | 4.5 Free
POF Free Dating App 4.0 | 3.0 Free
Badoo 4.0 | 4.5 Free
Zoosk 4.0 | 4.5 Free
JAUMO Flirt Chat 2.0 | 3.0 Free
Tinder 4.0 | 3.0 Free
LOVOO 2.0 | 4.5 Free
OkCupid Dating 2.0 | 3.0 Free

Match Dating 3.0 Free
OkCupid Dating 3.0 Free
Tinder 4.5 Free
JAUMO Flirt Chat 4.5 Free
Zoosk 3.0 Free
LOVOO 3.0 Free
Badoo 4.5 Free
POF Free Dating App 4.5 Free

FIGURE 2 | Screenshots of the simulated *list of apps* page. **(A)** For the lock group and warning system group. **(B)** For the control group and sound group.

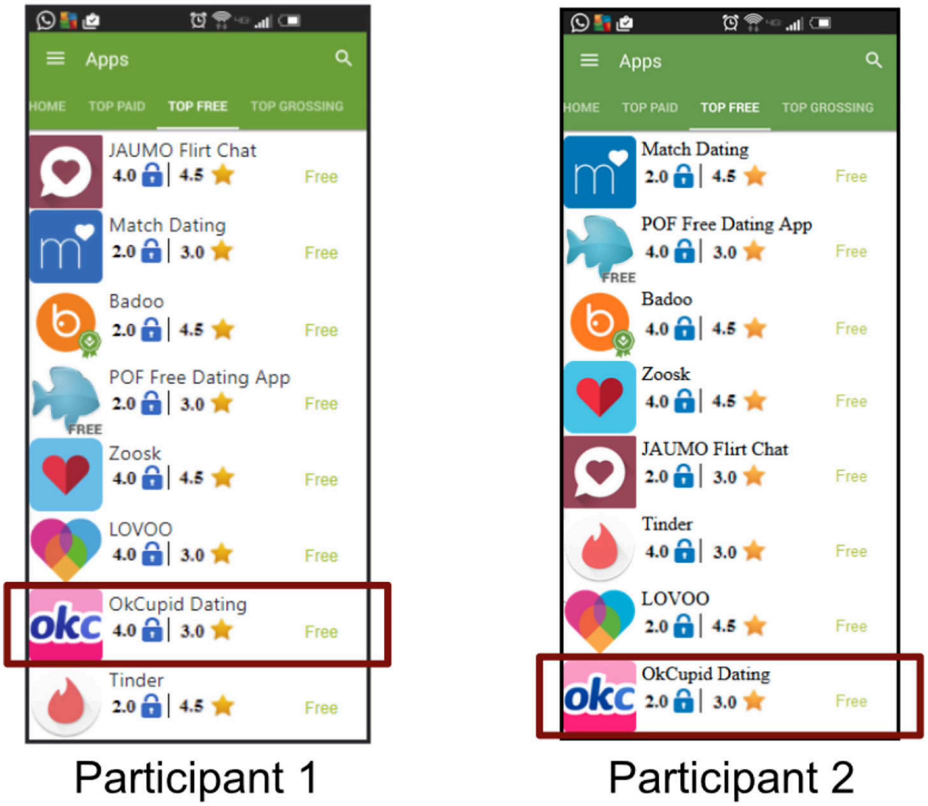


Match Dating
4.5 App Rating
100 THOUSAND Downloads
2.0 Privacy Rating
Click to view all Permissions

OkCupid Dating
4.5 App Rating
50 THOUSAND Downloads
Click to view all Permissions

FIGURE 3 | Screenshots of the simulated *app description* page. **(A)** For the lock group and warning system group. **(B)** For the control and sound group.

two remaining variables. In order to compute the values for PrivacyOverDownloadCount, we had to normalize the values for Privacy Rating and Download Count to be on the same scale. So the Download Count values 100 and 50 k would now be 4 and 2, respectively. We then compared the normalized values for Download Count and Privacy Rating against each other. If the Privacy Rating for a selected app was greater than the Download Count then PrivacyOverDownloadCount



Participant 1

JAUMO Flirt Chat
4.0 🔒 | 4.5 ★ Free

Match Dating
2.0 🔒 | 3.0 ★ Free

Badoo
2.0 🔒 | 4.5 ★ Free

POF Free Dating App
2.0 🔒 | 3.0 ★ Free

Zoosk
4.0 🔒 | 4.5 ★ Free

LOVOO
4.0 🔒 | 3.0 ★ Free

OkCupid Dating
4.0 🔒 | 3.0 ★ Free

Tinder
2.0 🔒 | 4.5 ★ Free

Participant 2

Match Dating
2.0 🔒 | 4.5 ★ Free

POF Free Dating App
4.0 🔒 | 3.0 ★ Free

Badoo
4.0 🔒 | 4.5 ★ Free

Zoosk
4.0 🔒 | 4.5 ★ Free

JAUMO Flirt Chat
2.0 🔒 | 3.0 ★ Free

Tinder
4.0 🔒 | 3.0 ★ Free

LOVOO
2.0 🔒 | 4.5 ★ Free

OkCupid Dating
2.0 🔒 | 3.0 ★ Free

FIGURE 4 | Screenshots of the *list of apps* page for two different participants highlighting randomization of attribute values.

was assigned to be 1, if Privacy Rating was equal to the Download Count then PrivacyOverDownloadCount was assigned to be 0, and if Privacy Rating was less than the Download Count the PrivacyOverDownloadCount was assigned to be −1. A similar approach was taken to compute the values for PrivacyOverAppRating.

Participants were asked to make 4 app choices per category. This was done to force a situation where it was necessary to make trade-offs between App Rating, Privacy Rating, and Download Count. If asked to make a single choice, participants could optimize across all three variables. By creating multiple choices, we obtain data on decisions where one factor must be chosen over another. In our analysis, we examine the ratio of the three variables to capture the results of these decisions. We choose categories where people tend to make multiple selections, particularly games. People engaged in online dating often also use multiple services (Valkenburg and Peter, 2007).

All three experimental variables were ordinal. For a given app, Privacy Rating(PR) was either 2 or 4, App Rating(AR) took on values 3 or 4.5 and Download Count(DC) was 50,000 or 100,000. We chose to go with higher values for App Rating when compared to the Privacy Rating because extensive past research showed that app ratings dominate choice in the absence of privacy indicators (Kelley et al., 2012; Rajivan and Camp, 2016). Additionally, participants would not want to install an app that is unusable and unwanted, even if it offered the highest privacy. We had adequate variance in app ratings to evaluate this using Generalized Estimating Equations (GEE). Using the values for the three experimental variables, we generated eight combinations of ratings: one app where all the variables had the lowest possible value, one app where all variables had the highest possible value, three apps where only one of the variables had the highest possible value, and three apps where at least two variables had the highest possible value. All eight combinations are listed below.

- Lowest possible values:
  {PR: 2, AR: 3 and DC: 50,000}
- Highest possible values:
  {PR: 4, AR: 4.5 and DC: 100,000}
- One variable with highest possible value:
  {PR: 4, AR: 3 and DC: 50,000}
  {PR: 2, AR: 4.5 and DC: 50,000}
  {PR: 2, AR: 3 and DC: 100,000}
- Two variables with highest possible values:
  {PR: 4, AR: 4.5 and DC: 50,000}
  {PR: 2, AR: 4.5 and DC: 100,000}
  {PR: 4, AR: 3 and DC: 100,000}

As mentioned in section 3.5, these combinations were randomly assigned to eight apps in each category. Requiring users to pick four out of the eight applications means that they cannot optimize all three experimental variables for all four app choices.

A participant can at most optimize two variables for two app choices, and for the remaining two choices, he/she can only optimize one experimental variable. This was done to force participants to prioritize one variable over the others.

We also created two example permissions manifests per app category such that one manifest represented over-permissions while the other represented least-permissions. The permission manifest that represented least-permissions was assigned to an app with a high Privacy Rating (4). Similarly, the permissions manifest that represented over-permissions was assigned to an app with a low Privacy Rating (2). This was done to provide internally consistent information. It also enabled privacy-aware participants in the Control Group to distinguish between privacy-persevering and privacy-invasive applications if they viewed the permissions.

In addition to the app choices, we also collected several implicit data measures from the experiment. These were permissions viewed, amount of time spent on choosing apps in each category, and the total time the participants took to complete the experiment.

4. EXPERIMENT AND PARTICIPANTS

The participants for this study were recruited from Amazon's Mechanical Turk (MTurk). Upon agreeing to participate in the study all participants were required to confirm that they owned an Android device. We achieved this by asking participants to visit an URL that provided them with a code only if they visited it using an Android device. Participants were required to have this code to continue with the study. We added this criteria for our study because we wanted to eliminate confounding factors originating from recruiting participants that don't use an Android device. Specifically, past work has shown that people using different platforms have different perceptions about the same app including privacy concerns (Ali et al., 2017; Mcilroy et al., 2017).

Next, all participants were provided with a simple set of instructions on how to use the interactive PlayStore simulator. The instructions were strictly mechanical, explaining that the participants had to select apps. After reading the instructions, the participants were allowed to progress to the simulated environment and make app choices. They were presented with two sets of app categories with eight apps in each category. After selecting the applications, participants answered demographic questions and questions for consistency checks. The order of categories, the order of apps under each category, and the ratings (Privacy Rating, App Rating, and Download Count) assigned to the apps were randomized for all participants. The categories were dating apps and puzzle apps.

Participants were asked to make four app choices in the order of their preference for each category, with the first choice being the most preferred and the fourth choice being the least preferred. Once the participants made all the necessary app choices, they were presented with queries about their app installation behavior, their computer literacy, and their demographics.

Reproductions of classic experiments have shown that the response of MTurk participants to priming and framing is consistent with participants in laboratory and field experiments (Horton et al., 2011). The use of MTurk is appropriate for this controlled study based on previous research and accepted practice (Horton et al., 2011; Casler et al., 2013; Chong et al., 2017). In methodologically validating related work conducted by Casler et al., participants were presented with four pairs of tools and they had to pick one tool from each pair to perform a task (Casler et al., 2013). While the in-lab participants were allowed to physically hold the tool, the Mturk participants only saw demonstrations of the tool being used. The researchers compared results from the laboratory study to that of the online simulation conducted on Mturk and found that the results were indistinguishable. In our work, participants perform the same actions to install or uninstall an application (the simulator replicated the interactions that users performed on the PlayStore) with a different mode of interaction (mouse vs. touch).

5. RESULTS

In the following, we begin with a rough summary and visualizations of the results. Then we provide a detailed statistical analysis.

5.1. Demographics

The study features four groups of subjects with three variables in each. Eighty participants were recruited for each experimental condition. In total, we enrolled 320 participants for our study. This was larger than the number required by power analysis by more than a factor two.

Out of the 320 participants, 17 participants were disqualified for providing contradicting answers to questions in the questionnaires. For example, the question "Do you review/read the permissions presented to you before you install an app from the Google PlayStore?" was asked twice. Participants that gave two different answers were disqualified. We also excluded all the results from the participants who took <3 min to complete the study. After applying the above mentioned exclusion criteria, we ended up with a total of 235 participants. These exclusion criteria were used to identify participants who only put minimal effort toward making app choices. We then repeated the analysis without excluding those who took <3 min; the results were stronger in that there were smaller *p*-values. However, here we include the analysis for the smaller sample as our initial study design included the 3-min-limit.

We applied a location qualification in MTurk to require all participants to be within the United States. Out of the 235 participants, 60.85% were male and 39.15% were female. The majority of the participants were 25–35 years old (50.21%). 23.4% of the participants were between 35 and 45 years old, 14.8% were 18–25 years old, and 11.4% were older than 45. We cannot argue that the sample was representative of the U.S. population as a whole. Other scholars have noted that MTurk use limits representativeness and participation (Stritch et al., 2016). Conversely, MTurk is widely used and thus these results can be compared to similar related work, with multiple studies indicating that MTurk is a reliable resource for high-quality data (Buhrmester et al., 2016).

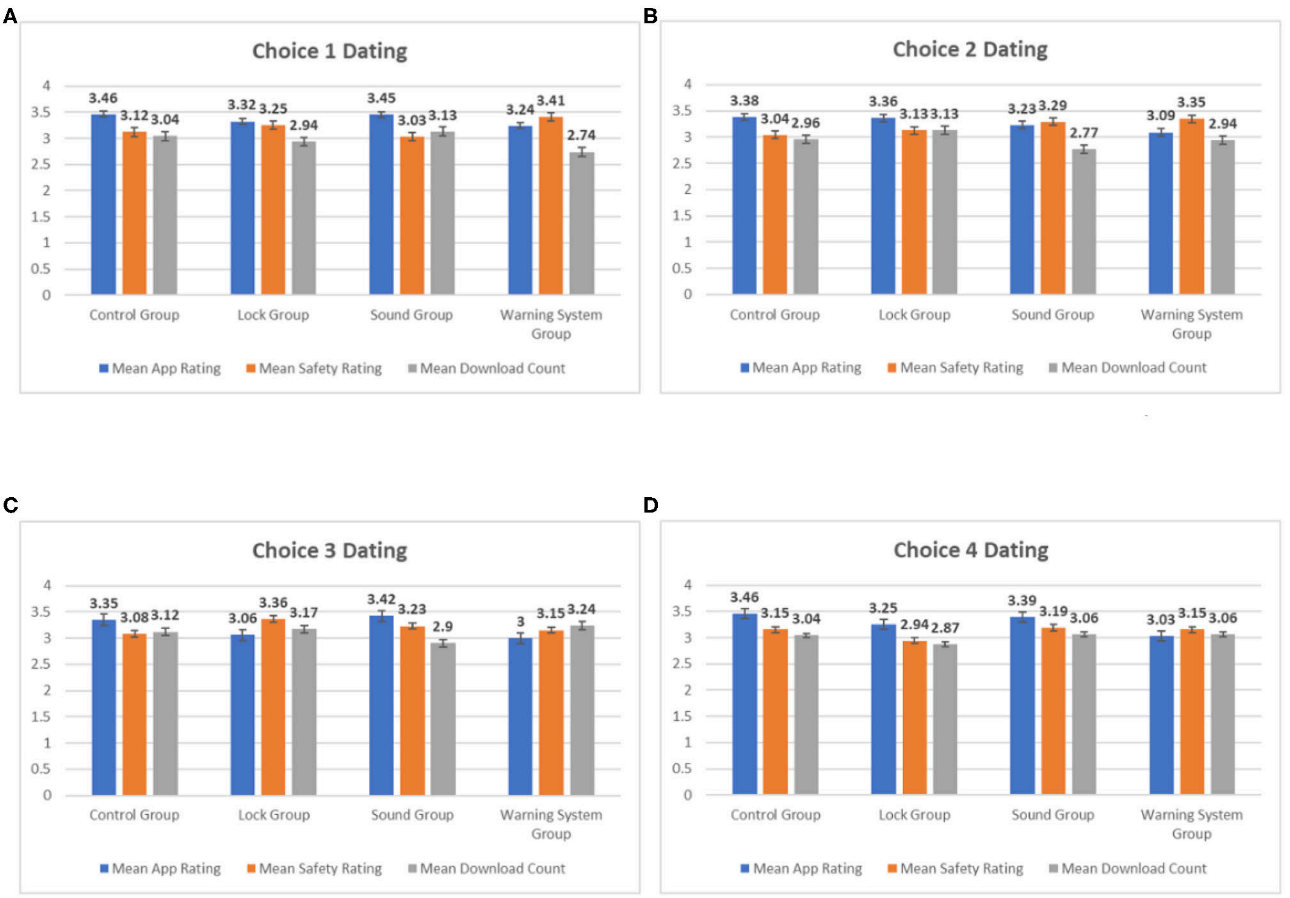


FIGURE 5 | Mean values for dating apps. **(A)** Means for app choice 1, **(B)** means for app choice 2, **(C)** means for app choice 3, and **(D)** means for app choice 4.

5.2. Basic Means Comparison

Figure 5 shows the histograms of mean App Rating, Privacy Rating, and Download Count for the four app choices in the dating category. As you can see in **Figure 5** the mean App Rating for all four choices in the control group is higher than the mean Privacy Rating and the mean Download Count. This indicates that participants in the Control Group were seeking a higher App Rating rather than maximizing Privacy Rating or Download Count. In contrast, the mean Privacy Rating is consistently higher than the mean App Rating and Download Count in the Warning System Group. Choice 3 is the only exception [Mean Download Count (3.24) is greater than Mean Privacy Rating (3.15)]. The mean Privacy Rating of the Warning System Group is higher than the mean Privacy Rating of the Control Group for the first three app choices. The mean Privacy Ratings for the fourth app choice are the same for both groups, but it is roughly equal to the App Rating. The Lock Group and the Sound Group also consistently showed a higher mean Privacy Rating when compared to the Control Group. Choice 1 is an exception for the Sound Group [Control Group (3.12) > Sound Group (3.03)] and choice 4 is an exception for the Lock Group [Control Group (3.15) > Sound Group (2.94)]. This shows that the Privacy Rating of the apps was higher when the participants were provided with the privacy cues. The trends are particularly clear in the Warning System Group.

Figure 6 shows the histograms of mean App Rating, Privacy Rating, and Download Count for the four app choices in the puzzles category. Similar to the dating apps, the mean App Rating for all four app choices in the Control Group is higher than the mean Privacy Rating and Download Count, indicating that participants' in the Control Group made their app choices that optimized App Rating. One other similarity is that the mean Privacy Rating for all four choices in the Warning System Group is higher than the mean App Rating and Download Count. This indicates that Privacy Rating had more influence on the app choices made by the participants in the Warning System Group when compared to the Control Group. This implication is strengthened by the fact that the mean Privacy Rating for the Warning System Group is higher than that of the Control Group for all four app choices. Also similar to the dating apps, the mean Privacy Ratings for the Lock Group and the Sound Group are higher than that of the Control Group for three out of four app choices [mean Privacy Rating Control Group (3.23) > mean Privacy Rating Sound Group (3.19) > mean Privacy Rating Lock Group (3.17) for Choice 3]. Privacy Rating appears

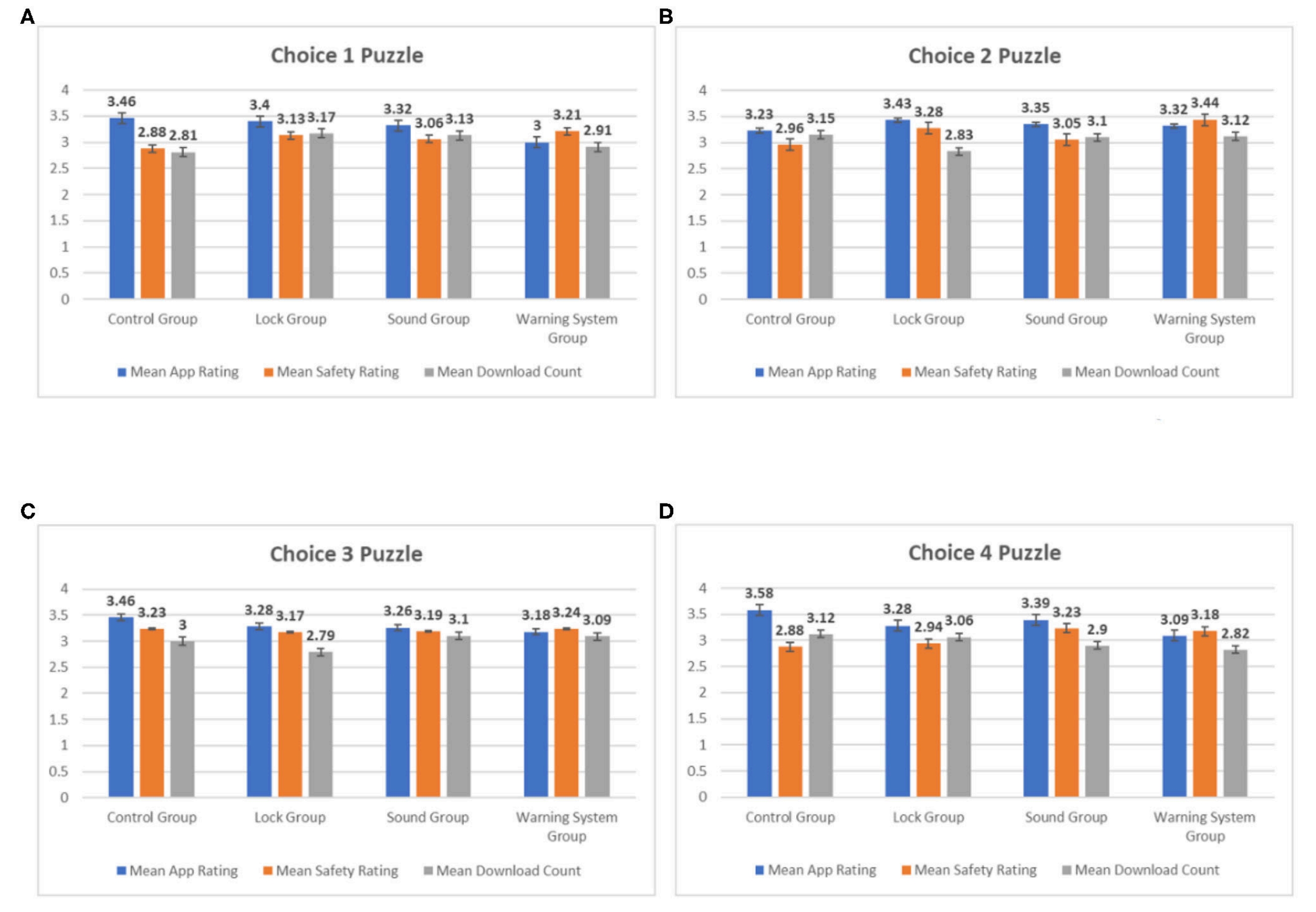


FIGURE 6 | Mean values for puzzle apps. **(A)** Means for app choice 1, **(B)** means for app choice 2, **(C)** means for app choice 3, and **(D)** means for app choice 4.

to have had more influence on app choices made by participants in groups with privacy cues when compared to the Control Group. Once again, this trend is most prominent in the Warning System Group.

5.3. Analysis

Typically, to determine if the difference between groups is statistically significant, a researcher would perform one-way Kruskal–Wallis and pairwise Mann–Whitney (pairwise comparison) tests for non-parametric data. These are commonly used to determine statistical differences between groups and are often requested by reviewers. However, in order for these tests to generate accurate results, the data must conform to certain assumptions. These assumptions are as follows:

- The dependent variable must be measured on an ordinal or continuous scale.
- The independent variable (in our case Groups) should have two or more categories.
- The observations must be independent (i.e., there should be no relationship between observations in each Group or between Groups).

Our study data violates one of the three assumptions. The recorded observations are not independent i.e., each participant makes four app installation choices which results in a dataset where the dependent variables (App Rating, Privacy Rating, PrivacyOverAppRating, PrivacyRatingOverDownloadCount) are correlated. If these correlations are not taken into account the results from the statistical analysis will not be valid and the results will be non-replicable. Therefore, to accurately determine the statistical differences between the control group and the experimental groups, we used Generalized Estimation Equations which requires no such assumptions.

5.3.1. Generalized Estimation Equations

Generalized Estimating Equations (GEE) are an extension of Generalized Linear Models and are commonly used to analyze correlated data that arises from repeated measurements (Hardin, 2005; Seago et al., 2006; Lee et al., 2007; Muth et al., 2016). In our case, the repeated measurements stem from each participant making four app installations in each category. A GEE analysis can evaluate the aggregate decisions to see if users in different groups behaved differently. GEE does not restrict the dependent variables to be continuous or require normal distribution. GEE aligns with our experimental goals and the resulting data.

When reporting the results from our analysis we provide both the p-value and the odds ratio. The p-value indicates the strength of the evidence against the null hypothesis and the odds ratio provides an effect size. The odds ratio represents the odds that an outcome will occur given a particular exposure compared to the odds of the outcome occurring in the absence of the exposure. In our case, the odds ratio is interpreted as follows. When the $\textit{Odds ratio} = 1$ this implies that the cues in the experimental group do not affect the outcome. When the $\textit{Odds ratio} > 1$ this indicates that participants in the experimental group are likely to have a higher value for the given dependent variable. When the $\textit{Odds ratio} < 1$ this indicates that the participants in the experimental group are likely to have a lower value for the given dependent variable.

5.3.2. Privacy Rating, App Rating, and Download Count

We performed GEE analysis on the collected data to see if Privacy Rating, App Rating, and Download Count were significantly different between the control group and the experimental groups. In this section, we report the results of our analysis.

The results for Privacy Rating are shown in **Table 2**. These results indicate that Privacy Rating is not significantly different from that of the Control Group for both Lock and Sound groups across the two app categories. For the Warning System Group, Privacy Rating is statistically significant for puzzle apps and marginally significant for dating apps. The odds ratio indicates that participants in the experimental groups are more likely to choose an app with a higher Privacy Rating when compared to that of the Control Group. Participants in the Warning System Group are 1.42 times more likely to select a dating app with a higher privacy rating and 1.76 times more likely to select a puzzle app with a higher privacy rating when compared to the Control group. The magnitude of the effect is clearly higher for puzzle apps when compared to dating applications. A visualization of this is provided in **Figure 7**.

The results from the analysis on App Rating can be found in **Table 3**. The results show that App Rating is statistically significant for the Warning System Groups across both app categories. For the Lock and Sound groups, the results are not statistically significant. The visualization for the odds ratio is shown in **Figure 8** shows that participants in the Control Group are more likely to select an app with a higher App Rating when compared to the Warning System Group. This effect was larger for dating apps when compared to puzzle apps.

Download Count was not found to be significant for all three experimental groups.

TABLE 2 | For the Warning System Group, the results are significant for puzzle apps and marginally significant for dating apps.

		p-values	Cohen's d
Warning system group	Dating apps	0.059	0.193
	Puzzle apps	0.001	0.312
Lock group	Dating apps	0.264	0.084
	Puzzle apps	0.063	0.159
Sound group	Dating apps	0.146	0.101
	Puzzle apps	0.063	0.154

For the remaining two experimental groups the results are not significant for both app categories.

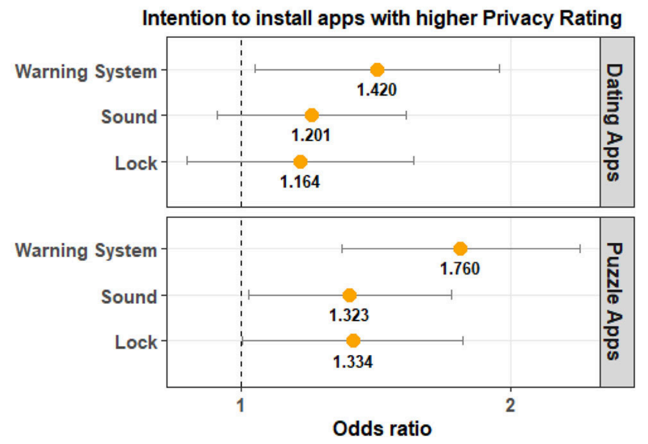


FIGURE 7 | The odds ratio (95% confidence interval) indicates that participants in the Warning System Group are more likely to select apps with a higher Privacy Rating compared to the Control Group.

TABLE 3 | The *p*-values show that App Rating is significantly different for the Warning System Group for both app categories.

		p-values	Cohen's d
Warning system group	Dating apps	*p* <0.001	−0.349
	Puzzle apps	0.002	−0.390
Lock group	Dating apps	0.074	−0.210
	Puzzle apps	0.285	−0.109
Sound group	Dating apps	0.465	−0.056
	Puzzle apps	0.179	−0.122

For the Lock Group and the Sound Group the results are not significant.

5.3.3. PrivacyOverAppRating and PrivacyOverDownloadCount

To understand the impact Privacy Rating had on the participants' app choices in comparison to App Rating and Download Count, we examined the ratio of Privacy Rating to App Rating as well as Privacy Rating to Download Count. To be more descriptive, we performed GEE analysis on the dependent variables PrivacyOverAppRating and PrivacyOverDownloadCount. As discussed in section 3.6, PrivacyOverAppRating tells us if the Privacy Rating for an installed app is greater than (1), equal to (0), or less than (−1) its App Rating. Similarly, PrivacyOverDownloadCount tells us if Privacy Rating for an installed app is greater than (1), equal to (0), or less than (−1) its Download count. A higher value for PrivacyOverAppRating or PrivacyOverDownloadCount indicates that participants attributed more weight to Privacy Rating at the time of app selection relative to App Rating and Download Count.

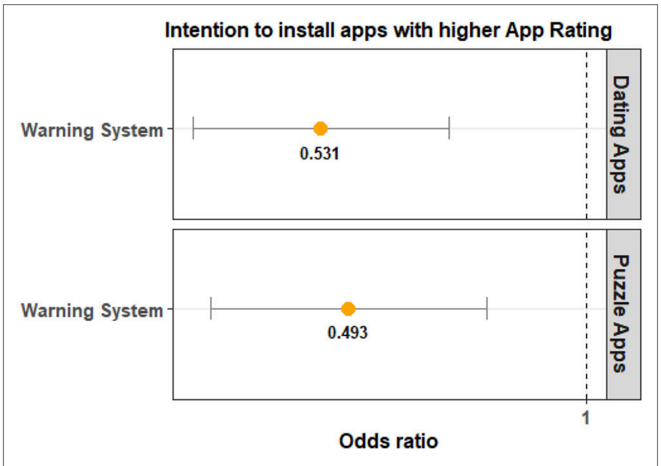


FIGURE 8 | The odds ratio (95% confidence interval) indicates that participants in the Warning System Group are more likely to select apps with a lower App Rating compared to the Control Group. Warning System is also consistently significant for both app categories.

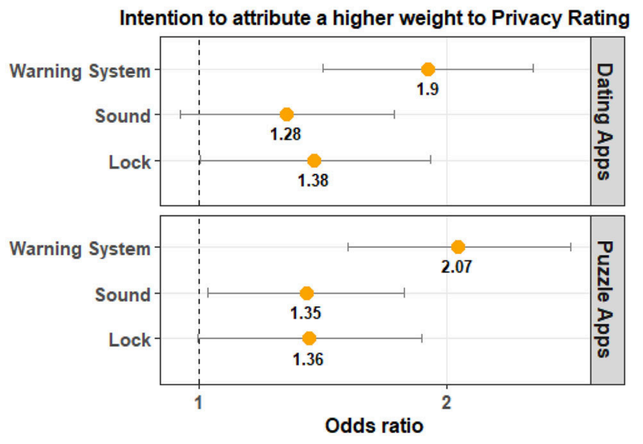


FIGURE 9 | The odds ratio (95% confidence interval) illustrates that participants in the Warning System Group are more likely to choose an app with a higher value for Privacy Rating relative to App Rating. Warning System is also consistently significant for both app categories and has a higher odds ratio compared to the Lock and Sound groups.

TABLE 4 | The *p*-values indicate that PrivacyOverAppRating is significantly different for the Warning System Group for both app categories.

		p-values	Cohen's d
Warning system group	Dating apps	*p* <0.001	0.354
	Puzzle apps	*p* <0.001	0.401
Lock group	Dating apps	0.063	0.178
	Puzzle apps	0.072	0.170
Sound group	Dating apps	0.150	0.136
	Puzzle apps	0.059	0.166

For the Sound Group, the results are marginally significant for puzzle apps.

TABLE 5 | For the Warning System Group, the results are significant for puzzle apps and marginally significant for dating apps.

		p-values	Cohen's d
Warning system group	Dating apps	0.059	0.157
	Puzzle apps	0.002	0.242
Lock group	Dating apps	0.329	0.058
	Puzzle apps	0.063	0.149
Sound group	Dating apps	0.074	0.127
	Puzzle apps	0.146	0.082

For the Lock and Sound groups the results are not significant.

Table 4 shows that the PrivacyOverAppRating is significantly different between the Control Group and the Warning System Group for both app categories. The odds ratio tells us that participants in the Warning System Group are more likely to have a higher PrivacyOverAppRating value when compared to the Control Group. This implies that Privacy Rating had a larger impact on the users' app choice when compared to App Rating. Once again the magnitude of the observed effect was larger for puzzle apps when compared to dating apps. The odds ratio and the visualization of the comparison can be seen in **Figure 9**.

For the Lock Group, the results were not statistically significant for both app categories. For the Sound Group, the results were not significant for dating apps and were marginally significant for puzzle apps. The odds ratio indicates that PrivacyOverAppRating is likely to be higher for the Lock Group and Sound Group. As shown in **Figure 9**, the magnitude of the effect is larger for the Warning System group.

The results for PrivacyOverDownloadCount are shown in **Table 5**. For the Warning System Group, the PrivacyOverDownloadCount is statistically significant for puzzle apps and marginally significant for dating apps. The results are not significant for the remaining two experimental groups. The odds ratio shows that the value of PrivacyOverDownloadCount is likely to be higher for the Warning System group. Similar to other instances, the magnitude of the effect is larger for the puzzle apps when compared to dating apps (see **Figure 10**).

To summarize, the Warning System Group is significantly more likely to have a higher value for both PrivacyOverAppRating and PrivacyOverDownloadCount than the Control Group. From this analysis, we can argue that participants with both visual and aural cues are more likely to make decisions reflecting a relatively higher attention to Privacy Rating.

5.3.4. App Installation Frequency

The efficacy of aural feedback may be a function of its novelty. Audio feedback in this work was implemented both as a form of priming, and for the negative sounds, as a warning. Excessive use of visual dialogs has desensitized people's awareness of security warnings on the web (Anderson et al., 2016; Vance et al., 2017). At the end of the survey, we asked participants how often they installed apps from Google's PlayStore. No one reported that they

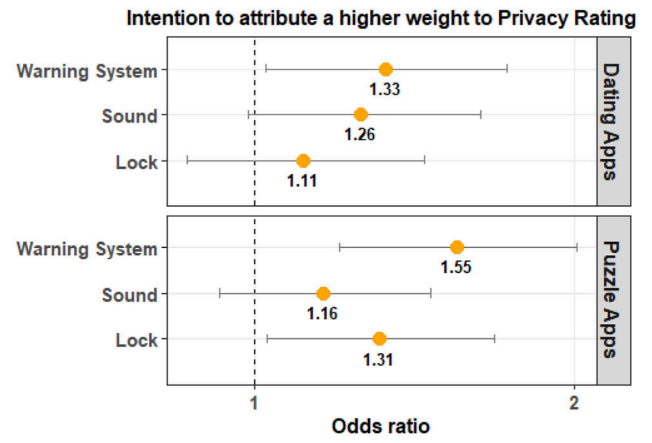


FIGURE 10 | The odds ratio (95% confidence interval) illustrates that participants in the Warning System Group are more likely to choose an app with a higher value for Privacy Rating relative to Download Count.

never installed apps from the app store. Respectively 15, 32, and 40% reported installing apps every other month, monthly, or weekly. A median user would see the warnings more often than once a month, and less often than once a week. The remaining participants reported that they installed apps on average every other day (9%), daily (2%), or more than once a day (also 2%). On average users would interact with the warnings every twenty-three days assuming thirty-day months. Habituation cannot be dismissed as a threat for all users, especially the 13% that would see the warning every other day or more. However, since 87% of the participants reported that they installed apps from the PlayStore *once a week* or *less often than once a week*, this indicates that for a large population habituation may be less of a concern. By definition, warning on first use only applies when a new app has been installed and is first run, app installation is an activity that does occur at roughly the same frequency as the first run or somewhat more often. Also note that, unlike warning dialogs, the specific audio feedback is unique and is not used by other computing devices. It is worth considering that our feedback does not interrupt the task flow. There is no dialog to close in this interaction, so this makes the communication potentially more acceptable it may also be easier to ignore over time.

5.3.5. Time to Decision

To determine if the addition of sound to the interaction was overwhelming, we compared the *time to decision* by participants in each condition. To further measure if the decision-making was burdensome, we conducted one-way ANOVA to test the differences of mean decision times between experimental groups. The differences in the means were not significant (p-value = 0.269). The mean times were 1.729, 1796, 1.760, and 1.859 for Control, Lock, Sound, and Warning System groups respectively. Previous work which compared different Internet panels for quality of data indicated that time to complete a survey was correlated with quality of data, and thus the decision to curtail participants by time to completion (Smith et al., 2016).

TABLE 6 | GEE results for Privacy Rating for data without the time filter with adjustments for multiple tests.

		p-values	Cohen's d
Warning system group	Dating apps	0.001	0.245
	Puzzle apps	$p < 0.001$	0.286
Lock group	Dating apps	0.010	0.148
	Puzzle apps	0.009	0.156
Sound group	Dating apps	0.002	0.189
	Puzzle apps	0.003	0.184

These results show that participants in all experimental groups made app choices that are significantly different from that of the control group.

6. DISCUSSION

As mentioned in section 5.3, the results show that people provided with both visual indicators and aural feedback are more likely to select apps with a higher Privacy Rating. This finding aligns with studies of warning systems offline, where information processing support impinges decision-making, and aural feedback is the most effective mode of communication at the time of exposure to a potential hazard.

In our study, we utilized attention check questions and time taken to install apps to identify and filter out participants who responded in an inattentive fashion. While attention check questions are known to be effective at identifying inattentive responses, response times were found to be unreliable for identifying inattentive responses (Downs et al., 2010; Gadiraju et al., 2015). The ineffectiveness of completion time as a filter could be due to the noise added by variability in computer load time, mouse maneuvering, and differences in cognitive processing time (Downs et al., 2010). Additionally, past research has shown that participants gaming the system use different strategies and take varying amounts of time (Gadiraju et al., 2015).

The decision to reconsider time as a variable was also influenced by the effect of attitudes on decision-making time (Fazio et al., 1989). Those familiar with the apps may have lower decision latency.

So it is not possible to separate the inattentive participants using completion time. As app installation time as a filter was a part of our initial study design, we reported results for participants who passed both attention checks. Since response time is now not considered a reliable method to filter out inattentive participants, here we report a subset of the results for all participants that passed the attention check questions without filtering out participants for app installation time. The complete results can be found in the Appendix.

Table 6 shows the results from the statistical analysis of data without the time filter for Privacy Rating. These results have been adjusted for multiple testing. The results show that Privacy Rating was significantly different from the Control Group for all

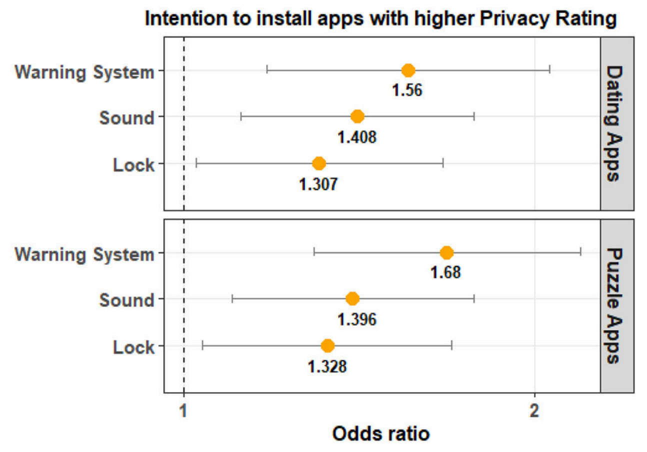


FIGURE 11 | The odds ratio (95% confidence interval) indicates that participants in all three experimental groups are more likely to select apps with a higher Privacy Rating. The effect size is larger for participants in the Warning System Group.

experimental groups. The odds ratio indicates that participants in all three experimental groups are more likely to select apps with a higher Privacy Rating. The effect size is larger for participants in the Warning System Group. This is illustrated in **Figure 11**.

The differences in results when decision time is not a filter indicate the potential for more research on how attention, decision time, and even distraction affect the efficacy of cues and warnings. These results show a clear significance for the Warning System across both categories. Sound is strongly significant for dating and puzzles; while Locks are similarly significant for both.

Under the most stringent analysis participants who were presented with only visual indicators or only audio feedback were not statistically different from the Control Group. This indicates that when people are presented with only visual indicators or audio feedback for privacy, they may not consistently make app choices that are privacy-preserving. This explain the inconsistent findings about privacy cues in previous work. This finding argues for more nuanced investigations on nudging privacy decision-making.

When the Privacy Rating was provided alongside the App Rating using only icons or only sound, we can not be entirely confident that participants' decisions were affected by the Privacy Rating. Without the audio feedback priming or warning participants to consider the Privacy Rating, they were less likely to pay attention to the visual cue. Conversely, when participants are provided with audio feedback but no visual indicator for Privacy Rating, then they may not be able to understand the implications of the audio feedback. As this is the first study on audio feedback in mobile resource warnings, more studies are needed to evaluate the efficacy of different sounds, or similar sounds with a different tone, pitch, and volume.

One possible reason for the disparity between the app choices for dating and puzzle apps could be that participants were more willing to share sensitive information with dating apps when compared to puzzle apps. It is clear why a dating app would require access to sensitive resources. For example, it is easy to understand that a dating app requires access to users' location to find people around them. But the same cannot be said about puzzle apps or game apps in general. For example, in a study conducted by Shklovski et al. participants felt deceived and expressed concerns when they learned about data collected by the Fruit Ninja app (Shklovski et al., 2014). In Lin et al. (2012), crowd-sourcing found that the acceptability of the same permissions varied across different apps.

Finally, regardless of cues, download count information was not significant in the app decision making process. Part of the reason could be that the download count values used for the experiment were not sufficiently different to influence app choices. Another reason could be that findings which indicate that download count dominates decision processes may have been observing a hidden variable (for example, the order of presentation or familiarity). We included the results for download counts in our paper because the lack of impact of download counts on participants' app choices is a significant finding even if it is only for relatively a smaller difference in download count. More research is needed to understand if larger variances in download count affect participants' app choices.

Our results indicate that participants who engaged with a multimedia warning system were more likely to make privacy-preserving app choices than those provided only with audio feedback or visual indicators. Consistent user awareness of privacy risks could have a significant cumulative effect on the entire mobile ecosystem. Given that one person's privacy choices impinge on the privacy of that person's contacts and potentially even those who share local area networks or physical location, a small but consistent improvement in mobile resource use by apps could have significant effects.

One further area of investigation is the relationship between fear and aural warnings. If the warnings create a fear response, this would be correlated with an increase in security behaviors (Johnston and Warkentin, 2010). In this case, the aural warning would have increased perceptions of privacy as a threat and decisions would be impinged by perceptions of self-efficacy and the efficacy of the response. Extensions of warnings research that includes protection-motivation theory and how behavior is impinged by fear could contribute to a more nuanced understanding of app selection behaviors (Herath and Rao, 2009).

Did these function as warnings to which users would become habituated or did they provide decision support that would remain valuable? Since past research has shown that people are less likely to become habituated to polymorphic warnings (Anderson et al., 2016; Vance et al., 2017) an evaluation of polymorphic aural warnings would be worthwhile. *In-situ* experiments that measure user behavior in the complex real world, without the focus here on isolating experimental variables in our controlled study, would be ideal. There is also a need for deep qualitative investigations of the privacy perspectives of end-users. Both *in-situ* evaluations and qualitative investigations should include participants with varying levels of privacy preferences and technical expertise.

7. CONCLUSION

Our experiment tested the efficacy of a visual cue, audio feedback, and a combination of these. We grounded this in usable security and were informed by heuristics from warning science. We provided padlocks as a visual privacy cue in the presence of a realistic distribution of apps both with and without audio feedback. We considered other options (such as haptic interactions and additional visual framing) for priming users for privacy. We chose audio feedback because haptic interactions are not clearly good or bad, and additional visual framing could be confounding or interrupt the task. Audio warnings also have been found to be effective in creating immediate awareness of physical hazards, and some effect was also seen here.

The results from our experiment showed that when participants were presented with both visual (positively framed padlocks) and aural indicators (cheers and jeers), they made app choices that included consideration of privacy ratings; i.e., individuals chose apps with higher privacy ratings over apps with higher app ratings. This was a significant change in behavior when compared to the Control Group, where participants made app decisions primarily based on app ratings. Reflecting on the body of previous research, those participants who saw only icons did not consistently make decisions that were correlated with higher app ratings. Hence, the inclusion of aggregate ratings and multimedia priming offers promise for supporting more informed decision making in online app stores. An added benefit of the approach we present here is that it could create competition or incentives to develop apps that are more conservative in terms of permission use. Currently, many apps are over-privileged perhaps in part because there is little to no marketplace benefit to minimizing permissions requests.

One of the limitations of our study is that we don't compare paid apps against free apps. However, we note that past work examined free Android apps and their paid counterparts, and showed that there is no evidence to support that the premium versions of the same app offered more privacy when compared to their counterparts (Han et al., 2019). Additionally, the current payment structures are based on monetization strategies, maintenance costs, and features not privacy (Ali et al., 2017).

These are promising results, yet additional research is indicated before the model of audio feedback and visual cues are accepted as ground truth. One area of future research is how to distinguish between two apps that have two different but close privacy ratings, for example between 2 and 2.3. This would suggest the use of a continuous sound variable, ranging from intensely negative to strongly positive. Such future work could be informed by a participatory design approach, as this offers promise in evaluating how different audio indicators may convey privacy information. This method may be particularly useful for the identification of continuous instead of discrete sound options. While this research was focused on detecting effects among the participants from the MTurk population, it is worth noting that screen readers do not consistently read nor report security indicators. Thus another avenue of future work would include the visually impaired.

Longitudinal investigations could determine if these effects are a result of a lack of familiarity or improved decision support.

DATA AVAILABILITY STATEMENT

Subject only to approval by the Institutional Review Board for data anonymization, the datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Indiana University IRB. The patients/participants provided their informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

SG: worked on building the play store simulator (experimental environment), contributed to study design, conducted the Mturk study, and performed data analysis. LC: principle investigator, and contributed to study design. OB: worked on building the play store simulator (experimental environment), and contributed to study design. All authors contributed to the article and approved the submitted version.

FUNDING

The authors declare that this study received funding from Cisco and Comcast. These funders were not involved in the study design, collection, analysis, interpretation of data, the writing of this article or the decision to submit it for publication.

ACKNOWLEDGMENTS

We would like to thank Dr. Apu Kapadia for his guidance and thank Jacob Abbott, Ahmed Alzahrani, and JangDong Seo for collaborating with us on the statistical analysis. This material was based upon work supported by the National Science Foundation under Grant No. CNS7111565375 CNS 1814518, CySP Grant DOD/H98230-19-1-0310, support from a Cisco Research Award No. 1377239, and funds from the Comcast Innovation Fund. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF, the US Government, Cisco, nor Comcast.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2020.02227/full#supplementary-material

REFERENCES

Acquisti, A., Brandimarte, L., and Loewenstein, G. (2015). Privacy and human behavior in the age of information. *Science* 347, 509–514. doi: 10.1126/science.aaa1465

Agarwal, Y., and Hall, M. (2013). "Protectmyprivacy: detecting and mitigating privacy leaks on iOS devices using crowdsourcing," in *Proceeding of the 11th Annual International Conference on Mobile Systems, Applications, and Services* (Taipei: ACM), 97–110. doi: 10.1145/2462456.2464460

Ali, M., Joorabchi, M. E., and Mesbah, A. (2017). "Same app, different app stores: a comparative study," in *2017 IEEE/ACM 4th International Conference on Mobile Software Engineering and Systems (MOBILESoft)* (Buenos Aires), 79–90. doi: 10.1109/MOBILESoft.2017.3

Anderson, B. B., Jenkins, J. L., Vance, A., Kirwan, C. B., and Eargle, D. (2016). Your memory is working against you: how eye tracking and memory explain habituation to security warnings. *Decis. Support Syst*. 92, 3–13. doi: 10.1016/j.dss.2016.09.010

Anderson, R., and Moore, T. (2009). Information security: where computer science, economics and psychology meet. *Philos. Trans. R. Soc. Lond. A* 367, 2717–2727. doi: 10.1098/rsta.2009.0027

Arzt, S., Rasthofer, S., Fritz, C., Bodden, E., Bartel, A., Klein, J., et al. (2014). "Flowdroid: Precise context, flow, field, object-sensitive and lifecycle-aware taint analysis for Android apps," in *Proceedings of the 35th ACM SIGPLAN Conference on Programming Language Design and Implementation, PLDI'14* (New York, NY: Association for Computing Machinery), 259–269. doi: 10.1145/2666356.2594299

Balebako, R., Schaub, F., Adjerid, I., Acquisti, A., and Cranor, L. (2015). "The impact of timing on the salience of smartphone app privacy notices," in *Proceedings of the 5th Annual ACM CCS Workshop on Security and Privacy in Smartphones and Mobile Devices, SPSM '15* (New York, NY: ACM), 63–74. doi: 10.1145/2808117.2808119

Benisch, M., Kelley, P. G., Sadeh, N., and Cranor, L. F. (2011). Capturing location-privacy preferences: quantifying accuracy and user-burden tradeoffs. *Pers. Ubiquit. Comput*. 15, 679–694. doi: 10.1007/s00779-010-0346-0

Benton, K., Camp, L. J., and Garg, V. (2013). "Studying the effectiveness of Android application permissions requests," in *2013 IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops)* (San Diego, CA), 291–296. doi: 10.1109/PerComW.2013.6529497

Beresford, A. R., Rice, A., Skehin, N., and Sohan, R. (2011). "Mockdroid: trading privacy for application functionality on smartphones," in *Proceedings of the 12th Workshop on Mobile Computing Systems and Applications* (Phoenix: ACM), 49–54. doi: 10.1145/2184489.2184500

Brustoloni, J. C., and Villamarín-Salomón, R. (2007). "Improving security decisions with polymorphic and audited dialogs," in *Proceedings of the 3rd Symposium on Usable Privacy and Security, SOUPS '07* (New York, NY: ACM), 76–85. doi: 10.1145/1280680.1280691

Buhrmester, M., Kwang, T., and Gosling, S. D. (2016). *Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality Data*? Washington, DC: American Psychological Association.

Byers, S., Cranor, L. F., Kormann, D., and McDaniel, P. (2004). "Searching for privacy: design and implementation of a p3p-enabled search engine," in *International Workshop on Privacy Enhancing Technologies* (Berlin, Heidelberg: Springer), 314–328. doi: 10.1007/11423409_20

Casler, K., Bickel, L., and Hackett, E. (2013). Separate but equal? A comparison of participants and data gathered via Amazon's MTurk, social media, and face-to-face behavioral testing. *Comput. Hum. Behav*. 29, 2156–2160. doi: 10.1016/j.chb.2013.05.009

Chen, J., Gates, C. S., Li, N., and Proctor, R. W. (2015). Influence of risk/safety information framing on Android app-installation decisions. *J. Cogn. Eng. Decis. Mak*. 9, 149–168. doi: 10.1177/1555343415570055

Choe, E. K., Jung, J., Lee, B., and Fisher, K. (2013). "Nudging people away from privacy-invasive mobile apps through visual framing," in *Human-Computer Interaction-INTERACT 2013*, eds P. Kotzé, G. Marsden, G. Lindgaard, J. Wesson, and M. Winckler (Berlin, Heidelberg: Springer), 74–91. doi: 10.1007/978-3-642-40477-1_5

Chong, I., Ge, H., Li, N., and Proctor, R. W. (2017). Influence of privacy priming and security framing on Android app selection. *Proc. Hum. Factors Ergon. Soc. Annu. Meet*. 61, 796–796. doi: 10.1177/1541931213601691

Costante, E., Hartog, J., and Petković, M. (2015). Understanding perceived trust to reduce regret. *Comput. Intell*. 31, 327–347. doi: 10.1111/coin.12025

Cranor, L. F., Guduru, P., and Arjula, M. (2006). User interfaces for privacy agents. *ACM Trans. Comput. Hum. Interact*. 13, 135–178. doi: 10.1145/1165734.1165735

Downs, J. S., Holbrook, M. B., Sheng, S., and Cranor, L. F. (2010). "Are your participants gaming the system? Screening mechanical Turk workers," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Atlanta, GA), 2399–2402. doi: 10.1145/1753326.1753688

Egele, M., Kruegel, C., Kirda, E., and Vigna, G. (2011). "PiOS: detecting privacy leaks in iOS applications," in *Network and Distributed Security Symposium* (San Diego, CA: ISOC).

Egelman, S., Cranor, L. F., and Hong, J. (2008). "You've been warned: an empirical study of the effectiveness of web browser phishing warnings," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '08* (New York, NY: ACM), 1065–1074. doi: 10.1145/1357054.1357219

Enck, W., Gilbert, P., Han, S., Tendulkar, V., Chun, B.-G., Cox, L. P., et al. (2014). Taintdroid: an information-flow tracking system for realtime privacy monitoring on smartphones. *ACM Trans. Comput. Syst*. 32:5. doi: 10.1145/2619091

Enck, W., Octeau, D., McDaniel, P., and Chaudhuri, S. (2011). "A study of Android application security," in *USENIX Security Symposium, Vol. 2* (San Francisco, CA), 2.

Fazio, R. H., Powell, M. C., and Williams, C. J. (1989). The role of attitude accessibility in the attitude-to-behavior process. *J. Consum. Res*. 16, 280–288. doi: 10.1086/209214

Felt, A. P., Chin, E., Hanna, S., Song, D., and Wagner, D. (2011). "Android permissions demystified," in *Proceedings of the 18th ACM Conference on Computer and Communications Security, CCS′11* (Chicago, IL; New York, NY: ACM), 627–638. doi: 10.1145/2046707.2046779

Felt, A. P., Ha, E., Egelman, S., Haney, A., Chin, E., and Wagner, D. (2012). "Android permissions: user attention, comprehension, and behavior," in *Proceedings of the Eighth Symposium on Usable Privacy and Security, SOUPS '12* (New York, NY: ACM), 3:1–3:14. doi: 10.1145/2335356.2335360

Gadiraju, U., Kawase, R., Dietze, S., and Demartini, G. (2015). "Understanding malicious behavior in crowdsourcing platforms: the case of online surveys," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul), 1631–1640. doi: 10.1145/2702123.2702443

Garg, V., and Camp, J. (2013). Heuristics and biases: implications for security design. *IEEE Technol. Soc. Mag*. 32, 73–79. doi: 10.1109/MTS.2013.2241294

Gates, C. S., Li, N., Peng, H., Sarma, B., Qi, Y., Potharaju, R., et al. (2014). Generating summary risk scores for mobile applications. *IEEE Trans. Depend. Secure Comput*. 11, 238–251. doi: 10.1109/TDSC.2014.2302293

Han, C., Reyes, I., Elazari Bar On, A., Reardon, J., Feal, S., Egelman, S., et al. (2019). "Do You Get What You Pay For? Comparing the Privacy Behaviors of Free vs. Paid Apps," in *Workshop on Technology and Consumer Protection (ConPro 2019), in conjunction with the 39th IEEE Symposium on Security and Privacy* (San Francisco, CA).

Han, J., Yan, Q., Gao, D., Zhou, J., and DENG, H. R. (2014). "Android or iOS for better privacy protection?," in *International Conference on Secure Knowledge Management in Big-Data Era (SKM 2014)* (Dubai).

Han, J., Yan, Q., Gao, D., Zhou, J., and Deng, R. H. (2013). "Comparing mobile privacy protection through cross-platform applications," in *Network and Distributed System Security Symposium* (Reston, VA: Internet Society).

Hardin, J. W. (2005). Generalized estimating equations (GEE). *Encyclop. Stat. Behav. Sci*. 2. doi: 10.1002/0470013192.bsa250

Helfinstein, S. M., Mumford, J. A., and Poldrack, R. A. (2015). If all your friends jumped off a bridge: the effect of others' actions on engagement in and recommendation of risky behaviors. *J. Exp. Psychol*. 144:12. doi: 10.1037/xge0000043

Herath, T., and Rao, H. R. (2009). Protection motivation and deterrence: a framework for security policy compliance in organisations. *Eur. J. Inform. Syst*. 18, 106–125. doi: 10.1057/ejis.2009.6

Horton, J. J., Rand, D. G., and Zeckhauser, R. J. (2011). The online laboratory: conducting experiments in a real labor market. *Exp. Econ*. 14, 399–425. doi: 10.1007/s10683-011-9273-9

Joeckel, S., Dogruel, L., and Bowman, N. D. (2017). The reliance on recognition and majority vote heuristics over privacy concerns when selecting smartphone apps among German and US consumers. *Inform. Commun. Soc*. 20, 621–636. doi: 10.1080/1369118X.2016.1202299

Johnston, A. C., and Warkentin, M. (2010). Fear appeals and information security behaviors: an empirical study. *MIS Quart*. 549–566. doi: 10.2307/25750691

Kelley, P. G., Consolvo, S., Cranor, L. F., Jung, J., Sadeh, N., and Wetherall, D. (2012). "A conundrum of permissions: installing applications on an Android smartphone," in *International Conference on Financial Cryptography and Data Security* (Berlin, Heidelberg: Springer), 68–79. doi: 10.1007/978-3-642-34638-5_6

Kelley, P. G., Cranor, L. F., and Sadeh, N. (2013). "Privacy as part of the app decision-making process," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Berlin, Heidelberg: ACM), 3393–3402. doi: 10.1145/2470654.2466466

Kelley, T., Amon, M. J., and Bertenthal, B. I. (2018). Statistical models for predicting threat detection from human behavior. *Front. Psychol*. 9:466. doi: 10.3389/fpsyg.2018.00466

Lee, J.-H., Herzog, T. A., Meade, C. D., Webb, M. S., and Brandon, T. H. (2007). The use of GEE for analyzing longitudinal binomial data: a primer using data from a tobacco intervention. *Addict. Behav*. 32, 187–193. doi: 10.1016/j.addbeh.2006.03.030

Lee, L., Egelman, S., Lee, J. H., and Wagner, D. (2015). Risk perceptions for wearable devices. *arXiv [Preprint]. arXiv:1504.05694*.

Liccardi, I., Pato, J., Weitzner, D. J., Abelson, H., and De Roure, D. (2014). "No technical understanding required: Helping users make informed choices about access to their personal data," in *Proceedings of the 11th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services, MOBIQUITOUS '14* (London), 140–150. doi: 10.4108/icst.mobiquitous.2014.258066

Lin, J., Amini, S., Hong, J. I., Sadeh, N., Lindqvist, J., and Zhang, J. (2012). "Expectation and purpose: understanding users' mental models of mobile app privacy through crowdsourcing," in *Proceedings of the 2012 ACM Conference on Ubiquitous Computing* (Pittsburgh, PA: ACM), 501–510. doi: 10.1145/2370216.2370290

Lin, J., Liu, B., Sadeh, N., and Hong, J. I. (2014). "Modeling users' mobile app privacy preferences: Restoring usability in a sea of permission settings," in *10th Symposium On Usable Privacy and Security* (Menlo Park, CA: $*SOUPS*$ *2014)*, 199–212.

Mcdonald, A. M., Reeder, R. W., Kelley, P. G., and Cranor, L. F. (2009). "A comparative study of online privacy policies and formats," in *International Symposium on Privacy Enhancing Technologies Symposium* (Berlin, Heidelberg: Springer), 37–55. doi: 10.1145/1572532.1572586

Mcilroy, S., Shang, W., Ali, N., and Hassan, A. E. (2017). User reviews of top mobile apps in Apple and Google app stores. *Commun. ACM* 60, 62–67. doi: 10.1145/3141771

Mileti, D., and Sorensen, J. (1990). *Communication of Emergency Public Warnings: A Social Science Perspective and State-of-the-Art Assessment*. Oak Ridge National Laboratories Technical Report, (ON: DE91004981). doi: 10.2172/6137387

Momenzadeh, B., Gopavaram, S., Das, S., Jean Camp, L. (2020). "Bayesian evaluation of user app choices in the presence of risk communication on android devices," In *International Symposium on Human Aspects of Information Security and Assurance* (Cham: Springer), 211–223.

Morton, A. (2014). "All my mates have got it, so it must be okay": constructing a richer understanding of privacy concerns-an exploratory focus group study," in *Reloading Data Protection*, eds S. utwirth, R. Leenes, and P. De Hert (Dordrecht: Springer), 259–298. doi: 10.1007/978-94-007-7540-4_13

Muth, C., Bales, K. L., Hinde, K., Maninger, N., Mendoza, S. P., and Ferrer, E. (2016). Alternative models for small samples in psychological research: applying linear mixed effects models and generalized estimating equations to repeated measures data. *Educ. Psychol. Measure*. 76, 64–87. doi: 10.1177/0013164415580432

Mylonas, A., Kastania, A., and Gritzalis, D. (2013). Delegate the smartphone user? Security awareness in smartphone platforms. *Comput. Secur*. 34, 47–66. doi: 10.1016/j.cose.2012.11.004

Nissenbaum, H. (1998). Protecting privacy in an information age: the problem of privacy in public. *Law Philos*. 17, 559–596. doi: 10.2307/3505189

Olejnik, K., Dacosta, I., Machado, J. S., Huguenin, K., Khan, M. E., and Hubaux, J.-P. (2017). "Smarper: Context-aware and automatic runtime-permissions for mobile devices," in *2017 IEEE Symposium on Security and Privacy (SP)* (San Jose, CA), 1058–1076. doi: 10.1109/SP.2017.25

Pan, E., Ren, J., Lindorfer, M., Wilson, C., and Choffnes, D. (2018). Panoptispy: characterizing audio and video exfiltration from Android applications. *Proc. Privacy Enhanc. Technol*. 2018, 33–50. doi: 10.1515/popets-2018-0030

Pandita, R., Xiao, X., Yang, W., Enck, W., and Xie, T. (2013). "Whyper: Towards automating risk assessment of mobile applications," in *USENIX Security Symposium* (Washington, DC), 527–542.

Papacharissi and Zizi (2010). Privacy as a luxury commodity. *First Monday*. 15:8. doi: 10.5210/fm.v15i8.3075

Patil, B. et al. (2016). Effective risk analysis and risk detection for Android apps. *Int. J. Comput. Appl*. 147. doi: 10.5120/ijca2016911130

Rajivan, P., and Camp, J. (2016). "Influence of privacy attitude and privacy cue framing on Android app choices," in *Twelfth Symposium on Usable Privacy and Security (SOUPS 2016)* (Denver, CO: USENIX Association).

Reyes, I., Wiesekera, P., Razaghpanah, A., Reardon, J., Vallina-Rodriguez, N., Egelman, S., et al. (2017). "Is our children's apps learning?" Automatically detecting COPPA violations," in *Workshop on Technology and Consumer Protection (ConPro 2017)* (San Jose, CA).

Schlegel, R., Kapadia, A., and Lee, A. J. (2011). "Eyeing your exposure: quantifying and controlling information sharing for improved privacy," in *Proceedings of the Seventh Symposium on Usable Privacy and Security* (Pittsburgh, PA: ACM), 14. doi: 10.1145/2078827.2078846

Seago, J. A., Spetz, J., Keane, D., and Grumbach, K. (2006). College students' perceptions of nursing: a GEE approach. *Nurs. Leadersh*. 19, 56–74. doi: 10.12927/cjnl.2006.18174

Shklovski, I., Mainwaring, S. D., Skúladóttir, H. H., and Borgthorsson, H. (2014). "Leakiness and creepiness in app space: perceptions of privacy and mobile app use," in *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems* (Toronto, ON: ACM), 2347–2356. doi: 10.1145/2556288.2557421

Shokri, R., Troncoso, C., Diaz, C., Freudiger, J., and Hubaux, J.-P. (2010). "Unraveling an old cloak: K-anonymity for location privacy," in *Proceedings of the 9th Annual ACM Workshop on Privacy in the Electronic Society, WPES'10* (New York, NY: Association for Computing Machinery), 115–118. doi: 10.1145/1866919.1866936

Smith, S. M., A.Roster, C., L.Golden, L., and S.Albaumb, G. (2016). A multi-group analysis of online survey respondent data quality: comparing a regular USA consumer panel to MTurk samples. *J. Bus. Res.* 69, 3139–3148. doi: 10.1016/j.jbusres.2015.12.002

Stritch, J. M., Pedersen, M. J., and Taggart, G. (2016). The opportunities and limitations of using mechanical Turk (MTURK). *Int. Publ. Manage.* 20, 489–511. doi: 10.1080/10967494.2016.1276493

Sunshine, J., Egelman, S., Almuhimedi, H., Atri, N., and Cranor, L. F. (2009). "Crying wolf: an empirical study of SSL warning effectiveness," in *Proceedings of the 18th Conference on USENIX Security Symposium, SSYM'09* (Berkeley, CA: USENIX Association), 399–416.

Tsai, J. Y., Egelman, S., Cranor, L., and Acquisti, A. (2011). The effect of online privacy information on purchasing behavior: an experimental study. *Inform. Syst. Res*. 22, 254–268. doi: 10.1287/isre.1090.0260

Valkenburg, P. M., and Peter, J. (2007). Who visits online dating sites? Exploring some characteristics of online daters. *CyberPsychol. Behav*. 10, 849–852. doi: 10.1089/cpb.2007.9941

Vance, A., Kirwan, B., Bjornn, D., Jenkins, J., and Anderson, B. B. (2017). "What do we really know about how habituation to warnings occurs over time? A longitudinal FMRI study of habituation and polymorphic warnings," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, CHI'17* (New York, NY: Association for Computing Machinery), 2215–2227. doi: 10.1145/3025453.3025896

Viscusi, W. K., and Zeckhauser, R. J. (1996). Hazard communication: warnings and risk. *Ann. Am. Acad. Polit. Soc. Sci.* 545, 106–115. doi: 10.1177/0002716296545001011

West, R. (2008). The psychology of security. *Commun. ACM* 51, 34–40. doi: 10.1145/1330311.1330320

Wijesekera, P., Baokar, A., Tsai, L., Reardon, J., Egelman, S., Wagner, D., et al. (2017). "The feasibility of dynamically granted permissions: aligning mobile privacy with user preferences," in *2017 IEEE Symposium on Security and Privacy (SP)* (San Jose, CA), 1077–1093. doi: 10.1109/SP.2017.51

Wogalter, M. S., DeJoy, D., and Laughery, K. R. (2005). *Warnings and Risk Communication*. Philadelphia, PA: CRC Press.

Xia, H., and Brustoloni, J. C. (2005). "Hardening web browsers against man-in-the-middle and eavesdropping attacks," in *Proceedings of the 14th International Conference on World Wide Web, WWW '05* (New York, NY: ACM), 489–498. doi: 10.1145/1060745.1060817

Zhou, Y., Zhang, X., Jiang, X., and Freeh, V. W. (2011). "Taming information-stealing smartphone applications (on Android)," in *Trust and Trustworthy Computing*, eds J. M. McCune, B. Balacheff, A. Perrig, A.-R. Sadeghi, A. Sasse, and Y. Beres (Berlin, Heidelberg: Springer), 93–107. doi: 10.1007/978-3-642-21599-5_7

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 21 October 2020
doi: 10.3389/fpsyg.2020.527886

Framing Effects on Online Security Behavior

Nuria Rodríguez-Priego[1,2], René van Bavel[1]*, José Vila[3] and Pam Briggs[4]

[1] Joint Research Centre, European Commission, Seville, Spain, [2] Departamento de Análisis Económico: Teoría Económica e Historia Económica, Universidad Autónoma de Madrid, Madrid, Spain, [3] Center for Research in Social and Economic Behavior (ERI-CES), Intelligent Data Analysis Laboratory (IDAL), University of Valencia, Valencia, Spain, [4] Department of Psychology, School of Life Sciences, Northumbria University, Newcastle upon Tyne, United Kingdom

We conducted an incentivized lab experiment examining the effect of gain vs. loss-framed warning messages on online security behavior. We measured the probability of suffering a cyberattack during the experiment as the result of five specific security behaviors: choosing a safe connection, providing minimum information during the sign-up process, choosing a strong password, choosing a trusted vendor, and logging-out. A loss-framed message led to more secure behavior during the experiment. The experiment also measured the effect of trusting beliefs and cybersecurity knowledge. Trusting beliefs had a negative effect on security behavior, while cybersecurity knowledge had a positive effect.

Keywords: cyber security, gain vs. loss frame, prospect theory, lab experiment, online behavior, nudge, threat assessment

OPEN ACCESS

Edited by:
Paul Watters,
Independent Researcher, Melbourne, Australia

Reviewed by:
Anton Kühberger,
University of Salzburg, Austria
Chris Baber,
University of Birmingham, United Kingdom

****Correspondence:***
René van Bavel
rene.van-bavel@ec.europa.eu

Specialty section:
This article was submitted to Cognition, a section of the journal Frontiers in Psychology

***Received:** 17 January 2020*
***Accepted:** 28 September 2020*
***Published:** 21 October 2020*

Citation:
Rodríguez-Priego N, van Bavel R, Vila J and Briggs P (2020) Framing Effects on Online Security Behavior. Front. Psychol. 11:527886. doi: 10.3389/fpsyg.2020.527886

INTRODUCTION

One of the many benefits of the digital transformation of markets is the ability for consumers to access a wide variety of stores and products from any device that connects to the Internet. However, this implies a growth in the complexity of consumer vulnerabilities, often exceeding regulatory efforts (Kucuk, 2016). Chief among these is cybercrime, a growing trend. The proportion of malicious URLs increased from 1 in 20 in 2016 to 1 in 13 in 2017 (SYMANTEC, 2018). In addition, threats in the use of mobile technology increased by 54 percent in 2017, compared to 2016, probably due to the rising use of these devices to access the Internet.

In order to remain secure online, consumers need to preserve their data confidentiality and integrity. They have to make cybersecurity decisions, respond to security-related messages and make adjustments to security-related settings that are not always easily understood (Payne and Edwards, 2008). Many consumers display limited cybersecurity knowledge and skills, despite having daily access to the Internet (Bennett et al., 2008; Bennett and Maton, 2010). Few are fully aware of the consequences of their online behavior, few see their behavior as risky and many fail to follow the recommendations and advice on safety given to them. All of which means that people end up behaving unsafely online, making them vulnerable to cyberattacks.

Such behavioral vulnerability means that people are often the weakest link in the cybersecurity chain (Sasse et al., 2001), which makes them a target. In 2017, 41% of ransomware attacks were against consumers (SYMANTEC, 2018); therefore, a better understanding of users' security behavior is relevant to tackling the problem of cybersecurity (Yan et al., 2018).

There are many actions consumers could take to increase their online security, including: running and updating antivirus software; using firewalls; not trusting in odd emails from unknown sources (Anderson and Agarwal, 2010); using strong passwords; logging out from sites; using trusted and secure connections, sites and services; providing the minimum amount of personal information needed; and being aware of physical surroundings (Coventry et al., 2014). Yet campaigns and training initiatives aimed at promoting such behaviors are often unsuccessful (Bada et al., 2019) and people generally ignore warnings (Junger et al., 2017), so more is being done to see how behavioral "nudges" might be designed to improve secure behavior and decision-making more directly.

To date a significant body of research has addressed behavioral issues in cybersecurity. For example, recent studies have shown that message framing can affect online shopping decisions (Cheng et al., 2014; Jin et al., 2017) and that privacy priming and security framing can generate safer decision-making around app selection (Chong et al., 2018) or change security incident reporting (Briggs et al., 2017). However, a significant issue with much of this previous research is that it has focused on *perceptions* of privacy and security risks (Miyazaki and Fernandez, 2001) or has over-relied upon *self-reported* past behaviors (Milne et al., 2009), or stated *behavioral intentions* (Anderson and Agarwal, 2010). This paper goes a step further and measures *observed behavior*. This is important, as studies of observed behavior drawn from both psychology and behavioral economics show human decision-making to be both flawed and biased. In part, this is because people are economic in their thinking and avoid processing details explicitly in order to make greater use of their automatic thinking and intuition (Milkman et al., 2009). By investigating actual consumer behaviors, we can understand more about the way such biases impact cybersecurity decision-making.

The present study contributes to a larger research initiative exploring the potential of behavioral insights to improving security behavior. It tests the effectiveness of two similar warning messages, designed to encourage consumers to behave more securely while shopping online, on a range of cybersecurity behaviors. In order to measure these behaviors, we created a lab environment designed to mimic the online shopping experience and provided them with a financial endowment to spend. We then gave participants either a message that focused on the positive outcomes resulting from behaving securely (i.e., a message that framed their behavior in terms of financial gain) or a message focused on negative outcomes resulting from not behaving securely (i.e., a message that framed their behavior in terms of financial loss). Critically, our messages reflected an actual financial gain or loss to the consumer. This is important to avoid adverse effects generated by giving supplemental warning messages that are not properly integrated into the task (Junger et al., 2017).

The rest of this article is structured as follows: section "Literature and Hypotheses" presents the literature review on framing effects and the hypotheses. Section "Materials and Methods" describes the methodology and the experimental procedure; section "Results" presents the results; and section "Conclusion" offers some conclusions.

LITERATURE AND HYPOTHESES

Individuals will react differently depending on how information is presented to them. In particular, when asked to choose between two options with the same expected value, people will be influenced by whether the outcome is framed as a gain (e.g., likelihood of winning) or as a loss (e.g., likelihood of losing). The frame does not alter the communicated content – it just presents it differently (Tversky and Kahneman, 1981; Druckman, 2001).

In their seminal work, Tversky and Kahneman (1981) presented experimental subjects with two options. One offered a certain outcome and the other offers an uncertain (i.e., risky) outcome. Both options had the same expected value (i.e., utility *x* probability). Options were framed in terms of gains or in terms of losses. Subjects tended to prefer the option of a certain (i.e., non-risky) gain over a risky gain. Conversely, they preferred options with an uncertain (i.e., risky) loss over a certain loss. In other words, people tend to avoid risks when facing the prospect of gains, but will seek risks to avoid prospective losses.

Loss aversion, or negativity bias, suggests people assign stronger values to negative feelings than to positive ones (Kahneman and Tversky, 1979; Rothman and Salovey, 1997). The impact and sensitivity of negative information, therefore, will be higher (Cacioppo et al., 1997; Baumann et al., 2019). For example, individuals display more distress when thinking about losing an amount of money, than the enthusiasm they exhibit for winning the same amount (McGraw et al., 2010). It follows that people will be more motivated to avoid losses than to pursue a gain of equal value (Rozin and Royzman, 2001; Vaish et al., 2008).

When an element of risk is introduced, the framing effect is more nuanced. In particular, in the gain frame, the risky prospect of having some losses is undesirable compared to the certain option of not having any losses. In the loss frame, the certain prospect of having some losses is undesirable compared to a risky prospect which could avoid losses altogether. Hence, in the gain frame people seek certainty and in the loss frame they accept risk (Zhang et al., 2017). In behavior change interventions, therefore, when individuals face a decision that involves a risk of obtaining an unpleasant outcome (e.g., cancer screening), loss-framed messages should be more effective. On the other hand, when the perceived risk of the unpleasant outcome is low, or when the outcome is pleasant (e.g., engaging in physical activity), a gain-framed message should work better (Rothman et al., 2006).

However, what can be expected of gain- and loss-framed messages in behavior change interventions more generally, where the element of risk is not present? The literature is ambiguous in this regard. On the one hand, interventions using a loss frame should be more effective in generating behavior change, simply because "losses loom larger than gains," as described above (see e.g., Hong et al., 2015). However, a number of sources in the literature argue that gain framing can also be effective as a longer-term intervention. In a meta-analysis of 93 disease prevention studies, gain-framed appeals were more persuasive than loss-framed appeals, although the difference was quite small and attributable to success in gain-framed messages promoting dental hygiene (O'Keefe and Jensen, 2008). Other sources report no significant differences overall, e.g., O'Keefe and Nan (2012) in a meta-analysis of vaccination behavior.

Other factors can mediate subjects' response to a framed message, such as the level of involvement with the issue, perceived self-efficacy, cultural background, the level of riskiness of the behavior itself, and socio-demographics (Maheswaran and Meyers-Levy, 1990; Banks et al., 1995; Rothman et al., 1999; Millar and Millar, 2000; Meyers-Levy and Maheswaran, 2004; Uskul et al., 2009; Lim and Noh, 2017). For example, in exploring the effects of interventions to reduce alcohol consumption, gain framed messages were more effective with those with low issue involvement, but loss-framed messages were found to be more effective in those with high issue involvement (de Graaf et al., 2015). In our own study, we ensured high issue involvement by making final payoff to the participants contingent upon their cybersecurity behavior and would therefore expect to see some cybersecurity benefits from a loss-framed message.

The Cybersecurity Context

Translating these findings to the cybersecurity context, we can see that to date, no studies have measured the direct behavioral impacts of a gain or loss framed cybersecurity message, although we can find one study that captures the advice a participant would offer to a fictional friend, following a gain-framed or loss-framed cybersecurity incident. Specifically, Rosoff et al. (2013) conducted a study in which people were presented with a set of scenarios in which they had fictional "prior experience" of a cybersecurity problem and were then asked to "advise a friend" as to the right action to take. Gain and loss framed messages were used to describe the potential outcome of a risky cyber choice with the gain-framed messages endorsing the safe, protective behaviors and the loss-framed messages warning of the consequences of risky action. For example, in a scenario about downloading music, the gain frame explained the actions to take for the friend to avoid the risk of acquiring a virus whereas the loss-frame highlighted the risk of them acquiring a virus. The authors found that the more the focus was on loss, the more likely participants were to make safer cybersecurity decisions. From this limited evidence of loss vs gain framing in the cybersecurity context, then, it would seem that losses do indeed loom larger than gains.

In our experiment, building upon the example above, we assume a loss-framed security message should be more effective in ensuring secure online behavior than a gain-framed message. We can also assume that, as the financial losses are real in our own paradigm, participants have high level of involvement, which would also contribute to loss-framing's effect. Based on these insights, we postulate the following hypothesis.

Hypothesis 1: The group exposed to the loss-framed message will show more secure online behavior than the group exposed to the gain-framed message.

We also consider other factors that could mediate the effect of the interventions tested. Trust is essential in the e-commerce environment as the process of buying online entails some risks, such as sharing personal information with an unknown seller. As a multidimensional construct, it refers to integrity, benevolence and predictability among other factors (McKnight et al., 2002; Gefen et al., 2003). Lack of trust toward an e-commerce seller may prevent users from buying online (Jarvenpaa et al., 1999; Grabner-Kräuter and Kaluscha, 2003; Gefen and Heart, 2006), conversely, trusting the vendor may facilitate online purchasing (McCole et al., 2010). This begs the question as to whether trust can lead to more reckless online behavior. It is an interesting issue and one which suggests an extension of the typical trust relationship in which vendor trust is a gateway to online purchasing. Here we ask whether vendor trust lead to riskier behavior all round. We would expect this to be the case, considering the antecedents of trust as discussed by Patrick et al. (2005), who point out how important trust is as a facilitator of social engineering attacks such as phishing, where familiarity with logos and trade names can lead consumers to erroneously place trust an online message. In this study, we wanted to assess whether trust in an online vendor can similarly create a "trust trap," effectively inducing a false sense of security that leads to a reduction of cybersecurity behaviors. Hence, we postulate that subjects who are more trusting will behave less securely as they may have confidence on vendor's goodwill and will not take the necessary steps to protect themselves. We measure *trusting beliefs* combining the scale developed by McKnight et al. (2002) and the one by Jarvenpaa et al. (1999). It provided a high internal consistency ($\alpha = 0.93$).

Hypothesis 2: Participants who exhibit higher levels of trust toward the vendor will show less secure online behavior than participants who exhibit lower levels of trust.

We also included a measure in our model related to *cybersecurity knowledge*, measured by asking our participants to assess a range of security-related behaviors (i.e., providing minimum information, connecting to a trusted site, logging out, etc. – see for example Coventry et al., 2014). We asked participants to rate the behaviors they thought could prevent them from suffering a cyberattack, using a 5-point Likert scale (1 = It won't reduce my risk at all; 5 = It will reduce my risk extremely). Internal consistency was tested through Cronbach's alpha and gave a high reliability of the scale ($\alpha = 0.90$). We expected higher levels of cybersecurity knowledge would lead to more secure behavior, either directly or through increased self-esteem (see e.g., Tang and Baker, 2016). Note that c*ybersecurity knowledge* was only measured in the post-purchase questionnaire to avoid participants being primed with this information during the experiment. We proposed the following hypothesis:

Hypothesis 3: Participants with a high level of cybersecurity knowledge will display more secure online behavior than participants with a lower level of knowledge.

MATERIALS AND METHODS

Experimental Procedure

We conducted a laboratory experiment with 120 participants, 60 per treatment[1]. The target population consisted of internet users who had purchased at least a product or a service online in the last 12 months. The participants were selected following a quota design for the sample of both treatments. The quotas were obtained from Eurostat's Annual Survey of Access and Usage of ICT in Households and Individuals 2013,

[1] This sample was extracted from a larger study with 600 participants testing the effect of different warning messages on security behavior (Rodríguez-Priego and van Bavel, 2016).

which established that internet users who purchased a good or service online in the previous 12 months in Spain were 51.7% men and that 40.6% of the Internet users were under 35 years of age. The sample was obtained from the subject pool managed by the laboratory of experimental economics of the ERI-CES (University of Valencia) with more than 25,000 volunteers. The recruitment system of the lab opened a call on its web page, only visible to those participants already registered in the database. Participants had to be actual members of the target population and answered filter questions to confirm this point. They were randomly assigned to experimental treatments until the representative quotas for age and gender were completed in each treatment. After that, no more participants of the age group or gender whose quota had been reached were allowed to register for the experiment. Ethical approval was granted by the Experimental Research Ethics Commission of the ERI-CES. Subjects were invited to the experimental laboratory and randomly assigned to a computer station. At the end of the experimental session, they received an anonymous payment in an enveloped identified only by the number of their station.

During the experiment, participants were asked to make several shopping decisions and were assigned an amount of money (an endowment). The incentive for participating in the experiment was divided in two. They received a fixed show-up fee for participating in the experiment and a variable fee that depended on the decisions they made during the online shopping process and on the random event of suffering a cyberattack. Subjects were told that they could receive a random cyberattack during the experiment. To increase the ecological validity of the experiment and to establish a decision environment similar to real-world Internet use, subjects were informed that the probability of being attacked would depend on the level of security of their online behavior. No specific information on which decisions actually increased or reduced this security level was provided to them. The use of performance-related incentives was relevant in this context to simulate the risks they might take when going online. In the lab, it is not possible to introduce a virus in their computer or make them feel the threat of a cyber-attack, since participants are not using their own computer. Specifically, the fact of suffering the random cyberattack would damage them by reducing their variable payoff at the end of the experiment. Consequently, if they behaved unsafely during the experiment, they could suffer a simulated cyberattack, and they would earn less money. On the contrary, if the behaved safely during the experiment, the probability of suffering a cyberattack would be the lowest and they would receive more money. This mechanism generated an incentive that is aligned with those in real-life situations: subjects aim to reduce the probability of suffering a random cyberattack.

After reading the instructions, and before the shopping experience began, participants filled a questionnaire with sociodemographic items. At the end of the purchase process, they completed a second questionnaire. It included questions related to trust in the e-commerce provider and cybersecurity knowledge.

In the experiment, participants had to buy a real product (a desktop wallpaper). They also had to make several security decisions, although – as mentioned earlier – they were not explicitly told about the potential consequences of each of these decisions. The intention was to let them behave as they would do in a non-experimental environment, where no feedback on security performance is available.

At the end of the experiment, participants had to answer a second questionnaire. After this post-experimental questionnaire, we provided participants with information on their accumulated probability of suffering a cyberattack due to their navigation. A random process then determined if they suffered the cyberattack or not (based on the above-mentioned probabilities). If they suffered the cyberattack, they would lose part of their variable endowment.

Experimental Conditions

We assigned participants to one of two experimental conditions showing different security messages. The experimental conditions presented a message focusing on the possible positive (i.e., gain-framed) and negative (i.e., loss-framed) outcomes related to their security behavior. Before they had to make any security-related decision, a message appeared as a pop-up in the center of the screen. Participants had to close the pop-up window to continue with the experiment. Then, the message moved to the upper part of the screen. The gain-framed message stated, "*Navigate safely. If you do, you could win de maximum final endowment.*" The loss-framed message stated, "*Navigate safely. If you don't, you could lose part of your final endowment.*"

The Dependent Variables

Probability of Suffering a Cyberattack

The first behavioral outcome measure in this study, taken from van Bavel et al. (2019), was the probability of suffering a cyberattack at the end of the experiment, which would reduce participants' variable payment. The probability was in the range of 5 to 65% and was calculated as a product of the five security decisions made during the experiment. From this minimum value of five percent, the selection of an unsecured connection, a non-trusted vendor or not logging out added 12 percentage points each to the probability of suffering a cyberattack. The sign-up process added another 24 percentage points in total. Lack of strength in the selected password added anywhere from zero percentage points (if the password met all seven six security criteria) to 12 points (if it met none). The non-compulsory information provided added between zero (if none of the items were answered) to 12 points (if subjects answered provided all of the items).

The probability of suffering the attack worked as an effective outcome measure of the security level of decisions made by the subjects: if they always proceeded in the most secure way this probability was kept at its minimum value (5%). On the other hand, if they selected the riskiest option at each step of the experiment, the probability reached its maximum value (65%). The maximum probability was set at a higher value than what could be expected when navigating well-known e-commerce sites in the real world. This was done to maintain a wide range of variation in the outcome measure. In addition, since participants did not actually know this value, it had no impact on their online behavior. Finally, although the probability of suffering a cyberattack was not related to the actual chances of suffering a cyberattack outside the experiment, the decisions that determined the probability were based on good security behavior in the real world (Coventry et al., 2014). This lack of prior information on how this variable is measured provided more ecological validity to the experiment. In real online purchases, consumers do not know in which percentage each of their actions is contributing to an increase in their probability of suffering a cyberattack.

Cybersecure Behavior

The second behavioral outcome measure was computed as the mean of the five security-related decisions that participants had to make during the experiment, described in more detail below: choosing a secure connection, choosing a strong password, providing minimum information in the sign-up process, choosing a trusted vendor and logging-out.

The decisions of choosing a secure connection, choosing a trusted vendor and logging-out were binary. The strength of the chosen password depended on seven rules that follow the usual parameters (Keith et al., 2007). Providing minimum information on the sign-up process meant completing as few of the eight optional cells requesting personal information. More information on these decisions is provided in the following subsection. Consequently, the variable *cybersecure_behavior* was computed as in Eq. (1).

$$\text{Cybersecurity\_behaviour} = \frac{\text{connection} + \frac{\text{password}}{7} + \frac{\text{sign}-\text{up}}{8} + \text{vendor} + \log - \text{out}}{5} \quad (1)$$

Security-Related Decisions

During the experiment, participants had to make five security-related decisions, which represented actions that users should take to protect themselves from cyberattacks (Coventry et al., 2014). We focused on decisions related to online purchasing processes that could be tested in an experiment. Participants had to make the decisions sequentially as follows:

Decision 1: Choosing a Secure Connection

The first action participants had to make was to connect to the experimental intranet. This was in fact a simulated intranet, with the only aim to examine participants′ security decisions. They had two options: they could choose to connect to the intranet through a secure or an unsecured connection. The secure connection forced the participants to wait 60 s and type a password provided on the screen. The purpose was to force them to make an extra effort if they wanted to behave securely. The next screen displayed a processing bar that charged during the connection process. Below the bar, participants could see a button that allowed them to change to an unsecured but immediate connection if they did not want to wait the entire minute. This possibility would let participants to change their mind, as in the real world.

The unsecured connection was an instant connection to the simulated intranet. Participants did not have to wait – the connection time was 0 s and it did not require any password. However, by choosing this option, participants increased their probability of suffering a cyberattack. The objective was to highlight the often intricate process that behaving safely online entails (as opposed to behaving unsafely). Choosing a secure option reflected the *compliance budget* that users weigh to make a decision (Beautement et al., 2009). The options (secure vs. unsecured) appeared randomly on the left or right-hand side of the screen to avoid location having an effect on participants' decisions.

After connecting to the intranet, participants could see the e-commerce website. It displayed the mock company name and logo, and a link to the terms and conditions. The link contained information about how the data would be managed, used and stored; the rights of the user; and copyright information. All this information complied with the European Data Protection Directive 95/46/EC. Participants had to accept the terms and conditions during the sign-up process by clicking the button "I agree to the Terms and Conditions".

The homepage was the gate for the subjects to start choosing products. When a subject clicked on a product, a detailed page for that product opened. On this page, the subject could click on the "buy" button to continue with the shopping process, or could go back to see any other products offered.

Decision 2: Choosing a Strong Password

Online consumers can prevent unauthorized individuals to exploit their password by creating a long password (Keith et al., 2007), or combining numbers and special characters with letters.

During the experiment, once subjects decided which product to buy, they had to register by creating a username and a password. We measured the level of password strength according to seven common security parameters, which included a minimum number of characters, lower case characters, upper case characters, numeric digit characters, and special characters, and a Boolean check whether password contained the username or email. Each of the seven criteria would increase the probability of suffering a cyberattack if not met.

Decision 3: Providing Minimum Information in the Sign-up Process

During the registration process, after choosing the username and password, participants were asked to provide some personal information. The information required to continue with the process was marked with an asterisk (name, surname, and email), but the remaining information (gender, age, phone number, address, zip code, city, region, and country) was optional. This is the usual kind of information requested in websites, which e-Commerce providers find useful for sending targeted advertising. The secure option was to disclose only the required information. Each of the eight non-compulsory items increased the probability of suffering a cyberattack. While the other four decisions reduced the risk of suffering a cyberattack, this measure went in the opposite direction: the higher the value meant the participant was behaving *less* securely. Therefore, when included in the outcome measure *cybersecure_behavior*, the "sign-up" variable was reversed. Admittedly, this variable had some limitations, as the veracity of the information provided in these non-compulsory items could not be guaranteed. In order to preserve anonymity, the personal data disclosed by participants was not recorded.

From the moment subjects registered until the end of the purchasing process, the top right-hand side of the screen displayed the text "Welcome" followed by their username, next to which was a button to log out of the e-commerce website.

Decision 4: Choosing a Trusted Vendor

Once subjects had completed the registration process, they had to select their choice of product (desktop wallpaper) between four possible options. Each of the products displayed a different picture, but the decision of choosing one of them was not relevant for the study, as it did not involve any secure or unsecure option. After that, participants had to choose between two vendors. Both vendors offered the same product, and were randomly ordered. The price offered by the first vendor for the product was zero. In this case, the link to download the product had no security signals (no image for an e-trusted site). The simulated link for this supplier was http (Hypertext Transfer Protocol). The second vendor offered the product for €2, but the link to download it was of the https (Hypertext Transfer Protocol Secure) type and appeared next to an image indicating it was an e-trusted site. Different prices depending on the security of the provider reflected how, in the real world, users can obtain products for free, but possibly compromising their security. If the participants chose the unsecured option (for free), they would increase the probability of suffering a cyberattack.

Decision 5: Logging Out

Once subjects had completed the purchasing process, a new screen displayed information about the cost of the purchased product and the amount remaining on their credit cards. A new button indicating "Next questionnaire" appeared at the bottom right-hand side of this screen. However, the secure option was to log out before continuing to the next questionnaire. Participants were not told explicitly to log out, although they were asked to exit the e-commerce website and complete the next questionnaire. If they did not log out, their probability of suffering a cyberattack at the end of the experiment increased.

RESULTS

In this section, we present the socio-demographic profile of participants in the sample and the ANCOVA model that tested the effect of the treatments, trust beliefs and knowledge on the probability of suffering a cyberattack.

Sociodemographic Information of the Sample

Quotas were applied by sex and age. Their value was fixed according to the profile of the internet users provided by the Annual Survey of Access and Usage of ICT in Households and Individuals in 2013, where 51.7% of Internet users were men and that 40.6% of the Internet users were under 35 years of age. Age ranged between 19 and 69 years. Sixty percent of participants were older than 32 and the mean age was 36.9 years. We provide further sociodemographic information on the educational level and employment status of the participants in **Table 1**.

Main Effects on the Probability of Suffering a Cyberattack

The mean probability of suffering a cyberattack during the experiment was higher in the gain-framed treatment ($M = 33.16$, SD = 10.04) than in the loss-framed treatment ($M = 28.43$, SD = 11.74; **Figure 1**). A two-tailed t-test comparing the means of the probability of suffering a cyberattack between the two treatments (gain vs. loss) showed a significant effect [$t(188) = 2.37$, $p = 0.019$]. A *post hoc* analysis using jStat with an alpha of 0.05 gave a power of 0.636. A loss-framed message appeared to be more effective in generating secure behavior, lending some support to Hypothesis 1.

We estimated a first regression model taking as dependent variable the probability of suffering a cyberattack. The explanatory variables were: (i) the treatments; (ii) cybersecurity knowledge, trusting beliefs; and (iii) the interactions between the treatments and the other explanatory variables. This first model showed no significant results for the interactions between the treatments and the other independent variables. In other words, the effect of the gain vs. loss-framed messages did not depend on cybersecurity knowledge or trusting beliefs.

TABLE 1 | Sociodemographic characteristics of participants[1].

Education level	**%**
No studies	0.83
Primary or lower secondary education	5.00
Upper secondary education and post-secondary, non-tertiary education	54.17
Bachelor's degree or equivalent	31.67
Postgraduate degree	4.17
PhD	4.17
Employment status	**%**
Self-employed	3.33
Employed by a public or private institution	33.33
Unemployed	24.17
Homemaker	1.67
Student	35.00
Disabled	0.00
Retired	2.50

[1] *This table provides information on education level and employment status of the sample. Further information on gender and age is included in the subsection Sociodemographic Information of the Sample.*

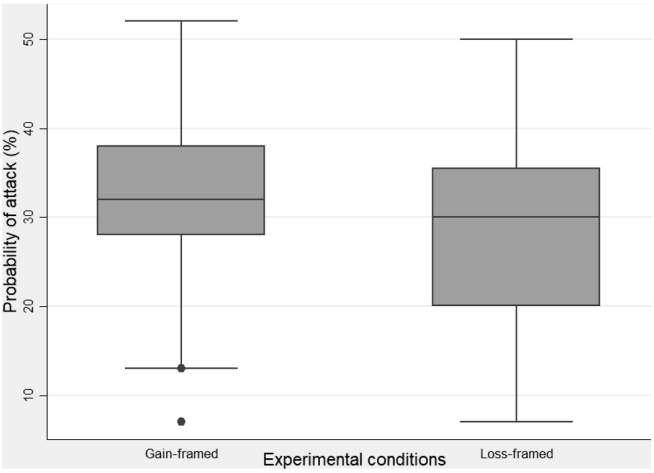


FIGURE 1 | Box-plot of the probability of suffering a cyberattack by experimental group.

Table 2 provides the estimation of the final model. It shows that the loss-framed message significantly decreased the probability of cyberattack compared to the gain-framed message [$t(116) = -2.36$, p-value = 0.020]. The estimated values of the coefficients show that a loss-framed message reduces the probability of suffering a cyberattack by 4.61%. This result confirms support for Hypothesis 1.

Second, *trusting beliefs* had a significant effect on the dependent variable [$t(116) = 2.15$, p-value = 0.034]. Participants who placed higher levels of trust in the vendor showed less secure behavior during the experiment. Hypothesis 2 is also supported.

Finally, knowledge of cybersecurity risks affected the probability of suffering a cyberattack in an inverse relationship (more knowledge meant less likelihood of an attack) [$t(116) = -2.13$, $p = 0.036$]. Hypothesis 3 is also supported.

Tables 3–7 show participants' behavior in each of the five decisions they had to make during the experiment, by experimental treatment. Regarding the first behavior (**Table 3**), all subjects decided to choose a secure connection over the unsecured one, no matter the framing of the message. Perhaps, at this early stage of the process, all subjects were concerned with navigating securely, as they had just read the security message that appeared in the center of the screen. After closing the pop-up, the message would only appear in the upper part of the screen during the rest of the experiment.

The second decision was to choose a password (**Table 4**). As mentioned before, password strength was measured according to seven common security parameters. Each of the seven criteria would increase the probability of suffering a cyberattack if not met. Results show that subjects in the loss-framed message condition met at least three of the seven criteria, and one of them met all criteria. In the gain-framed condition, three participants met fewer than three criteria and none of them met the seven criteria.

Table 5 shows the quantity of information that subjects provided during the sign-up process. There were eight non-compulsory items included in the sign-up information. Results show that 6.67% of subjects in the gain-framed condition provided no information apart from the compulsory, compared to 11.67% in the loss-framed condition.

The fourth decision was to choose between a trusted vs. untrusted vendor (**Table 6**). Here, 30% of participants in the

TABLE 2 | Estimated coefficients of the final model for the probability of suffering a cyberattack.

	Estimate	Std. Error	*t*-value	Pr(>|t|)
Loss-framed[1]	−4.61	1.95	−2.36	0.020
Knowledge[2]	−3.41	1.60	−2.13	0.036
Trusting beliefs[3]	2.92	1.36	2.15	0.034
Cons	−35.83	6.74	5.32	0.000

[1] *The gain-framed condition was taken as baseline for the data analysis.*
[2] *The variable Knowledge was estimated as a mean of the values obtained in each of the 10 items that comprised the Knowledge Scale. This scale is provided in the **Supplementary Table A2**. Each of the items were measured in a 5-point Likert scale.*
[3] *The variable Trusting beliefs was estimated as a mean of the values obtained in each of the 10 items that comprised the Trusting Beliefs Scale. This scale is provided in the **Supplementary Table A1**. Each of the items were measured in a 5-point Likert scale.*

TABLE 3 | Decision 1 – choosing a secure connection by treatment[1].

Treatment	Connection security		
	Unsecured	Secure	Total
Gain-framed[2]	0	60	60
%	0	100.00	100.00
Loss-framed[2]	0	60	60
%[3]	0	100.00	100.00
Total	0	120	120

[1] *Decision 1 was binary. It takes the value of 1 for choosing a secure connection, and 0 for choosing an unsecure connection.*
[2] *Values for gain-framed and loss-framed are given in absolute terms.*

TABLE 4 | Decision 2 – choosing a strong password by treatment[1].

Treatment	Password strength [1–7]							
	1	2	3	4	5	6	7	Total
Gain-framed[2]	1	2	16	17	23	1	0	60
%	1.67	3.33	26.67	28.33	38.33	1.67	0.00	100.00
Loss-framed[2]	0	0	16	20	17	6	1	60
%	0.00	0.00	26.67	33.33	28.33	10.00	1.67	100.00
Total	1	2	32	37	40	7	1	120

χ^2(6, N = 120) = 8.7147 Pr = 0.190.

[1] *Values for decision 2 ranged between 0 and 7 depending on the number of criteria that participants met for password strength. All of the subjects met at least 1 criteria.*

[2] *Values for gain-framed and loss-framed are given in absolute terms.*

TABLE 5 | Decision 3 – providing minimum information in the sign-up by treatment[1].

Treatment	Information provided in the sign-up [1–8]									
	0	1	2	3	4	5	6	7	8	Total
Gain-framed[2]	4	1	5	2	0	1	5	3	39	60
%	6.67	1.67	8.33	3.33	0.00	1.67	8.33	5.00	65.00	100.00
Loss-framed[2]	7	3	6	1	2	1	0	4	36	60
%	11.67	5.00	10.00	1.67	3.33	1.67	0.00	6.67	60.00	100.00
Total	11	4	3	2	2	2	5	7	75	120

χ^2 (8, N = 120) = 9.5053 Pr = 0.301.

[1] *Values for decision 3 ranged between 0 and 8 depending on the number of non-compulsory cells that participants filled in when registering in the e-commerce website.*

[2] *Values for gain-framed and loss-framed are given in absolute terms.*

TABLE 6 | Decision 4 – choosing a trusted vendor by treatment[1].

Treatment	Trusted vendor		
	Untrusted	Trusted	Total
Gain-framed[2]	18	42	60
%	30.00	70.00	100.00
Loss-framed[2]	10	50	60
%	16.67	83.33	100.00
Total	28	92	120

χ^2 (1, N = 120) = 2.981, p = 0.084.

[1] *Decision 4 was binary. It takes the value of 1 for choosing a trusted vendor, and 0 for choosing an untrusted vendor.*

[2] *Values for gain-framed and loss-framed are given in absolute terms.*

gain-framed treatment decided to choose the untrusted vendor, compared to a 16.67% of subjects who visualized the loss-framed message.

The last decision was to log-out or stay connected at the end of the purchase process (**Table 7**). The amount of participants who chose the secure option (i.e., logging-out) was a 15% higher in the loss-framed condition than in the gain-framed one. Finally, although we found differences between both treatments in some of the individual security-related decisions, none of them was statistically significant.

TABLE 7 | Decision 5 – logging out by treatment[1].

Treatment	Logging out		
	Stay connected	Log out	Total
Gain-framed[2]	48	12	60
%	80.00	20.00	100.00
Loss-framed[2]	39	21	60
%	65.00	35.00	100.00
Total	87	33	120

χ^2 (1, N = 120) = 3.3856 Pr = 0.066

[1] *Decision 5 was binary. It takes the value of 1 for logging-out after the purchase, and 0 for staying connected.*

[2] *Values for gain-framed and loss-framed are given in absolute terms.*

TABLE 8 | Estimated coefficients of the final model for cybersecure behavior.

	Estimate	Std. error	*t*-value	Pr(>|t|)
Loss-framed[1]	0.07	0.03	2.46	0.015
Knowledge[2]	0.05	0.03	2.16	0.033
Trusting beliefs[3]	−0.05	0.02	−2.24	0.027
Cons	0.50	0.11	4.59	0.000

[1] *The gain-framed condition was taken as baseline for the data analysis.*

[2] *The variable Knowledge was estimated as a mean of the values obtained in each of the 10 items that comprised the Knowledge Scale. This scale is provided in the **Supplementary Table A2**. Each of the items were measured in a 5-point Likert scale.*

[3] *The variable Trusting beliefs was estimated as a mean of the values obtained in each of the 10 items that comprised the Trusting Beliefs Scale. This scale is provided in the **Supplementary Table A1**. Each of the items were measured in a 5-point Likert scale.*

Main Effects on Cybersecure Behavior

Table 8 provides the estimated coefficients of the model for the dependent variable *cybersecure_behavior*. It shows that the loss-framed message significantly increased cybersecure compared to the gain-framed message [$t(116) = 2.46$, p-value = 0.015]. A *post hoc* analysis using jStat with an alpha of 0.05 gave a power of 0.653. The estimated values of the coefficients show that a loss-framed message increases cybersecure behavior, which supports Hypothesis 1.

Trusting beliefs had also a significant effect on the dependent variable [$t(116) = -2.24$, p-value = 0.027], which confirms Hypothesis 2. Participants who placed higher levels of trust in the vendor showed less secure behavior during the experiment.

Third, *knowledge* of cybersecurity risks influenced positively cybersecure behavior, providing support for Hypothesis 3 [$t(116) = 2.16$, p-value = 0.033].

CONCLUSION

In this research, we examined the effect of security messages on Internet users' behavior during an online shopping process. Our first hypothesis was that, compared to gain-framed messages, loss-framed messages would be more effective in ensuring participants behaved securely during this process. The findings support this hypothesis.

This paper then makes a contribution by extending work on loss aversion bias, where individuals assign stronger values to negative feelings than to positive ones (Kahneman and Tversky, 1979; Rozin and Royzman, 2001; Ert and Erev, 2008; Vaish et al., 2008; McGraw et al., 2010), and shows its relevance to the cybersecurity context.

A number of recent studies, including Junger et al. (2017), suggest the presence of threat information can backfire if it takes the form of a general warning, yet in our study threat or loss information was effective. Two aspects of our loss-framing might be relevant here.

Firstly, our loss message was tied explicitly to a financial loss outcome (i.e., it did not simply cite some kind of general threat). This means our result is in line with the idea that messages focused on the negative consequences of non-compliance are more persuasive (Cacioppo et al., 1997) when participants are more involved, i.e., more motivated to change. In our case, participants stood to lose money if they behaved insecurely and so motivation (or involvement) was high (cf. de Graaf et al., 2015). Our findings also demonstrate that the "loss looms larger" message does apply to cybersecurity behavior and is not limited to behavioral intentions [as with the Rosoff et al. (2013) study].

Secondly, our loss message was yoked to a behavioral nudge to navigate safely (i.e., we told consumers what they needed to do to avoid loss). Therefore, our intervention was aligned to recent findings that show that threat (or loss) appeals in isolation fail, but they can be effective when presented in conjunction with coping messages that direct consumer behavior (van Bavel et al., 2019).

With regard to trusting beliefs, subjects who trusted the vendor more performed worse on the experiment, meaning that they made decisions that entailed more security risks, ending with a higher probability of suffering a cyberattack. This result supports our second hypothesis and ties in with the literature on phishing and other forms of social engineering wherein trust in a known vendor is explicitly used to overcome defensive behaviors (Patrick et al., 2005). Consequently, trusting beliefs and their influence on users' performance as the weakest link in this wider cybersecurity chain is an issue that should be further investigated.

It should not be surprising that trust is an issue in this space. Firstly, we know that trust in an e-commerce vendor not only increases click-through intention, but also decreases malware risk perception (Ogbanufe and Kim, 2018). Secondly, and more importantly, we have seen the "weaponisation" of trust, with the huge rise in cybersecurity attacks that draw on social engineering principles to create an illusion of trust. Consumers are often led to believe that communication is with a "trusted" party, when in fact some imitation of that trusted party occurs (e.g., in phishing attacks). Trust, when exploited in this way, has negative implications for both genuine vendors and consumers and it is interesting to explore the kinds of "nudges" that might make people less willing to trust in a superficially familiar message or website (e.g., Moody et al., 2017; Nicholson et al., 2017).

The results regarding the effect of knowledge about cybersecurity support our third hypothesis. Subjects with a higher level of agreement that the listed security actions would prevent them from being attacked behaved more secure during the experiment. We can extract from this that subjects who have a clear concern of what secure behavior means may perform better when making security decisions – a finding again in keeping with recent work on the role of promoting "coping interventions" as part of cybersecurity protection (e.g., Tsai et al., 2016; Jansen and van Schaik, 2017; van Bavel et al., 2019).

Our findings from the questionnaire confirm that consumers' trust makes them vulnerable and that knowing what secure behavior is improves security decisions. Based on our experimental findings, however, we would contend that a fear-arousal behavioral component that describes a meaningful loss, but that also describes the way to avoid that loss, could be effective as a cybersecurity intervention.

DATA AVAILABILITY STATEMENT

The datasets generated for this study can be found in the Mendeley Data Repository (Rodriguez-Priego, Nuria (2020), "Framing effects on online security behavior", Mendeley Data, V2, doi: 10.17632/sp6cyrfvrv.2).

ETHICS STATEMENT

Ethical approval was granted by the Experimental Research Ethics Commission of the ERI-CES from the University of Valencia. All participants provided informed consent.

AUTHOR CONTRIBUTIONS

All authors contributed equally to the work.

FUNDING

The study was part of the European Commission's project Behavioral Insights on Cybersecurity (JRC/SVQ/2014/J.3/0039/RC-AMI).

ACKNOWLEDGMENTS

We are grateful to Ioannis Maghiros, Xavier Troussard, and Fabiana Scapolo at the Joint Research Centre for their continued support.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2020.527886/full#supplementary-material

REFERENCES

Anderson, C. L., and Agarwal, R. (2010). Practicing safe computing: a multimedia empirical examination of home computer user security behavioural intentions. *MIS Q.* 34, 613–643. doi: 10.2307/25750694

Bada, M., Sasse, M. A., and Nurse, J. R. (2019). Cyber security awareness campaigns: why do they fail to change behaviour?. *arXiv* [preprint]. Available online at: https://arxiv.org/abs/1901.02672 (accessed July 14, 2020).

Banks, S. M., Salovey, P., Greener, S., Rothman, A. J., Moyer, A., Beauvais, J., et al. (1995). The effects of message framing on mammography utilization. *Health Psychol.* 14:178. doi: 10.1037/0278-6133.14.2.178

Baumann, F., Benndorf, V., and Friese, M. (2019). Loss-induced emotions and criminal behavior: an experimental analysis. *J. Econ. Behav. Organ.* 159, 134–145. doi: 10.1016/j.jebo.2019.01.020

Beautement, A., Sasse, M. A., and Wonham, M. (2009). "August. The compliance budget: managing security behaviour in organisations," in *Proceedings of the 2008 Workshop on New Security Paradigms* (New York, NY: ACM), 47–58.

Bennett, S., and Maton, K. (2010). Beyond the 'digital natives' debate: towards a more nuanced understanding of students' technology experiences. *J. Comput. Assist. Learn.* 26, 321–331. doi: 10.1111/j.1365-2729.2010.00360.x

Bennett, S., Maton, K., and Kervin, L. (2008). The 'digital natives' debate: a critical review of the evidence. *Br. J. Educ. Technol.* 39, 775–786. doi: 10.1111/j.1467-8535.2007.00793.x

Briggs, P., Jeske, D., and Coventry, L. (2017). "The design of messages to improve cybersecurity incident reporting," in *Proceedings of the International Conference on Human Aspects of Information Security, Privacy, and Trust* (Cham: Springer), 3–13. doi: 10.1007/978-3-319-58460-7_1

Cacioppo, J. T., Gardner, W. L., and Berntson, G. G. (1997). Beyond bipolar conceptualizations and measures: the case of attitudes and evaluative space. *Pers. Soc. Psychol. Rev.* 1, 3–25. doi: 10.1207/s15327957pspr0101_2

Cheng, F., Wu, C., and Lin, H. (2014). Reducing the influence of framing on internet consumers' decisions: the role of elaboration. *Comput. Hum. Behav.* 37, 56–63. doi: 10.1016/j.chb.2014.04.015

Chong, I., Ge, H., Li, N., and Proctor, R. W. (2018). Influence of privacy priming and security framing on mobile app selection. *Comput. Security* 78, 143–154. doi: 10.1016/j.cose.2018.06.005

Coventry, L., Briggs, P., Jeske, D., and van Moorsel, A. (2014). "Scene: a structured means for creating and evaluating behavioural nudges in a cyber security environment," in *Proceedings of the International Conference of Design, User Experience, and Usability* (Cham: Springer), 229–239. doi: 10.1007/978-3-319-07668-3_23

de Graaf, A., van den Putte, B., and de Bruijn, G. (2015). Effects of issue involvement and framing of a responsible drinking message on attitudes, intentions, and behaviour. *J. Health Commun.* 20, 989–994. doi: 10.1080/10810730.2015.1018623

Druckman, J. N. (2001). Evaluating framing effects. *J. Econ. Psychol.* 22, 91–101. doi: 10.1016/s0167-4870(00)00032-5

Ert, E., and Erev, I. (2008). The rejection of attractive gambles, loss aversion, and the lemon avoidance heuristic. *J. Econ. Psychol.* 29, 715–723. doi: 10.1016/j.joep.2007.06.003

Gefen, D., and Heart, T. (2006). On the need to include national culture as a central issue in e-commerce trust beliefs. *J. Glob. Inform. Manag.* 14, 1–30. doi: 10.4018/jgim.2006100101

Gefen, D., Karahanna, E., and Straub, D. W. (2003). Trust and TAM in online shopping: an integrated model. *MIS Q.* 27, 51–90. doi: 10.2307/30036519

Grabner-Kräuter, S., and Kaluscha, E. A. (2003). Empirical research in on-line trust: a review and critical assessment. *Int. J. Hum. Comput. Stud.* 58, 783–812. doi: 10.1016/s1071-5819(03)00043-0

Hong, F., Hossain, T., and List, J. A. (2015). Framing manipulations in contests: a natural field experiment. *J. Econ. Behav. Organ.* 118, 372–382. doi: 10.1016/j.jebo.2015.02.014

Jansen, J., and van Schaik, P. (2017). Comparing three models to explain precautionary online behavioural intentions. *Inform. Comput. Security* 25, 165–180. doi: 10.1108/ics-03-2017-0018

Jarvenpaa, S. L., Tractinsky, N., and Saarinen, L. (1999). Consumer trust in an Internet store: a cross-cultural validation. *J. Comput. Med. Commun.* 5:JCMC526.

Jin, J., Zhang, W., and Chen, M. (2017). How consumers are affected by product descriptions in online shopping: event-related potentials evidence of the attribute framing effect. *Neurosci. Res.* 125, 21–28. doi: 10.1016/j.neures.2017.07.006

Junger, M., Montoya, L., and Overink, F. (2017). Priming and warnings are not effective to prevent social engineering attacks. *Comput. Hum. Behav.* 66, 75–87. doi: 10.1016/j.chb.2016.09.012

Kahneman, D., and Tversky, A. (1979). Prospect theory: an analysis of decision under risk. *Econometrica* 47, 263–292. doi: 10.2307/1914185

Keith, M., Shao, B., and Steinbart, P. J. (2007). The usability of passphrases for authentication: an empirical field study. *Int. J. Hum. Comput. Stud.* 65, 17–28. doi: 10.1016/j.ijhcs.2006.08.005

Kucuk, S. U. (2016). Consumerism in the digital age. *J. Consum. Affairs* 50, 515–538. doi: 10.1111/joca.12101

Lim, J. S., and Noh, G. (2017). Effects of gain-versus loss-framed performance feedback on the use of fitness apps: mediating role of exercise self-efficacy and outcome expectations of exercise. *Comput. Hum. Behav.* 77, 249–257. doi: 10.1016/j.chb.2017.09.006

Maheswaran, D., and Meyers-Levy, J. (1990). The influence of message framing and issue involvement. *J. Mark. Res.* 27, 361–367. doi: 10.2307/3172593

McCole, P., Ramsey, E., and Williams, J. (2010). Trust considerations on attitudes towards online purchasing: the moderating effect of privacy and security concerns. *J. Bus. Res.* 63, 1018–1024. doi: 10.1016/j.jbusres.2009.02.025

McGraw, A. P., Larsen, J. T., Kahneman, D., and Schkade, D. (2010). Comparing gains and losses. *Psychol. Sci.* 21, 1438–1445. doi: 10.1177/0956797610381504

McKnight, D. H., Choudhury, V., and Kacmar, C. (2002). Developing and validating trust measures for e-commerce: an integrative typology. *Inform. Syst. Res.* 13, 334–359. doi: 10.1287/isre.13.3.334.81

Meyers-Levy, J., and Maheswaran, D. (2004). Exploring message framing outcomes when systematic, heuristic, or both types of processing occur. *J. Consum. Psychol.* 14, 159–167. doi: 10.1207/s15327663jcp1401%262_18

Milkman, K. L., Chugh, D., and Bazerman, M. H. (2009). How can decision making be improved? *Perspect. Psychol. Sci.* 4, 379–383.

Millar, M. G., and Millar, K. U. (2000). Promoting safe driving behaviours: the influence of message framing and issue involvement. *J. Appl. Soc. Psychol.* 30, 853–856. doi: 10.1111/j.1559-1816.2000.tb02827.x

Milne, G. R., Labrecque, L. I., and Cromer, C. (2009). Toward an understanding of the online consumer's risky behavior and protection practices. *J. Consum. Affairs* 43, 449–473. doi: 10.1111/j.1745-6606.2009.01148.x

Miyazaki, A. D., and Fernandez, A. (2001). Consumer perceptions of privacy and security risks for online shopping. *J. Consum. Affairs* 35, 27–44. doi: 10.1111/j.1745-6606.2001.tb00101.x

Moody, G. D., Galletta, D. F., and Dunn, B. K. (2017). Which phish get caught? An exploratory study of individuals' susceptibility to phishing. *Eur. J. Inform. Syst.* 26, 564–584. doi: 10.1057/s41303-017-0058-x

Nicholson, J., Coventry, L., and Briggs, P. (2017). "Can we fight social engineering attacks by social means? Assessing social salience as a means to improve phish detection," in *Proceedings of the Thirteenth Symposium on Usable Privacy and Security ({SOUPS} 2017)* (Berkeley, CA: USENIX Association), 285–298.

Ogbanufe, O., and Kim, D. J. (2018). "Just how risky is it anyway?" The role of risk perception and trust on click-through intention. *Inform. Syst. Manag.* 35, 182–200. doi: 10.1080/10580530.2018.1477292

O'Keefe, D. J., and Jensen, J. D. (2008). Do loss-framed persuasive messages engender greater message processing than do gain-framed messages? A meta-analytic review. *Commun. Stud.* 59, 51–67. doi: 10.1080/10510970701849388

O'Keefe, D. J., and Nan, X. (2012). The relative persuasiveness of gain-and loss-framed messages for promoting vaccination: a meta-analytic review. *Health Commun.* 27, 776–783. doi: 10.1080/10410236.2011.640974

Patrick, A. S., Briggs, P., and Marsh, S. (2005). Designing systems that people will trust. *Security Usabil.* 1, 75–99.

Payne, B. D., and Edwards, W. K. (2008). A brief introduction to usable security. *IEEE Internet Comput.* 12, 13–21. doi: 10.1109/mic.2008.50

Rodríguez-Priego, N., and van Bavel, R. (2016). *The Effect of Warning Messages on Secure Behaviour Online: Results from a Lab Experiment*. JRC Technical Reports, EUR 28154. Brussels: European Union.

Rosoff, H., Cui, J., and John, R. S. (2013). Heuristics and biases in cyber security dilemmas. *Environ. Syst. Decis.* 33, 517–529. doi: 10.1007/s10669-013-9473-2

Rothman, A. J., Bartels, R. D., Wlaschin, J., and Salovey, P. (2006). The strategic use of gain-and loss-framed messages to promote healthy behaviour: how theory can inform practice. *J. Commun.* 56(Suppl._1), S202–S220.

Rothman, A. J., Martino, S. C., Bedell, B. T., Detweiler, J. B., and Salovey, P. (1999). The systematic influence of gain- and loss-framed messages on interest in and use of different types of health behaviour. *Pers. Soc. Psychol. Bull.* 25, 1355–1369. doi: 10.1177/0146167299259003

Rothman, A. J., and Salovey, P. (1997). Shaping perceptions to motivate healthy behaviour: the role of message framing. *Psychol. Bull.* 121, 3–19. doi: 10.1037/0033-2909.121.1.3

Rozin, P., and Royzman, E. B. (2001). Negativity bias, negativity dominance, and contagion. *Pers. Soc. Psychol. Rev.* 5, 296–320. doi: 10.1207/s15327957pspr0504_2

Sasse, M. A., Brostoff, S., and Weirich, D. (2001). Transforming the 'weakest link'—a human/computer interaction approach to usable and effective security. *BT Technol. J.* 19, 122–131.

SYMANTEC (2018). *Internet Security Threats Report*. Available At: http://www.symantec.com/threatreport/last (accessed July 06, 2019).

Tang, N., and Baker, A. (2016). Self-esteem, financial knowledge and financial behaviour. *J. Econ. Psychol.* 54, 164–176. doi: 10.1016/j.joep.2016.04.005

Tsai, H. Y. S., Jiang, M., Alhabash, S., LaRose, R., Rifon, N. J., and Cotten, S. R. (2016). Understanding online safety behaviors: a protection motivation theory perspective. *Comput. Security* 59, 138–150. doi: 10.1016/j.cose.2016.02.009

Tversky, A., and Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science* 211, 453–458. doi: 10.1126/science.7455683

Uskul, A. K., Sherman, D. K., and Fitzgibbon, J. (2009). The cultural congruency effect: culture, regulatory focus, and the effectiveness of gain-vs. loss-framed health messages. *J. Exp. Soc. Psychol.* 45, 535–541. doi: 10.1016/j.jesp.2008.12.005

Vaish, A., Grossmann, T., and Woodward, A. (2008). Not all emotions are created equal: the negativity bias in social-emotional development. *Psychol. Bull.* 134, 383–403. doi: 10.1037/0033-2909.134.3.383

van Bavel, R., Rodríguez-Priego, N., Vila, J., and Briggs, P. (2019). Using protection motivation theory in the design of nudges to improve online security behaviour. *Int. J. Hum. Comput. Stud.* 123, 29–39. doi: 10.1016/j.ijhcs.2018.11.003

Yan, Z., Robertson, T., Yan, R., Park, S. Y., Bordoff, S., Chen, Q., et al. (2018). Finding the weakest links in the weakest link: how well do undergraduate students make cybersecurity judgment? *Comput. Hum. Behav.* 84, 375–382. doi: 10.1016/j.chb.2018.02.019

Zhang, X., Liu, Y., Chen, X., Shang, X., and Liu, Y. (2017). Decisions for others are less risk-averse in the gain frame and less risk-seeking in the loss frame than decisions for the self. *Front. Psychol.* 8:1601. doi: 10.3389/fpsyg.2017.0160

Disclaimer: The views expressed in this article are purely those of the authors and may not in any circumstances be regarded as stating an official position of the European Commission.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 23 October 2020
doi: 10.3389/fpsyg.2020.535605

Identifiability, Risk, and Information Credibility in Discussions on Moral/Ethical Violation Topics on Chinese Social Networking Sites

Xi Chen, Chenli Huang and Yi Cheng*

School of Business Administration and Tourism Management, Yunnan University, Kunming, China

OPEN ACCESS

Edited by:
Richard Keith Wortley,
University College London,
United Kingdom

Reviewed by:
Hu Jian,
Jiangxi University of Finance
and Economics, China
Pei Huang,
Beijing University of Posts
and Telecommunications (BUPT),
China
Ivan Wen,
University of Hawai'i at Mānoa,
United States

****Correspondence:***
Xi Chen
13610491@qq.com

Specialty section:
This article was submitted to
Cognition,
a section of the journal
Frontiers in Psychology

Received: *01 April 2020*
Accepted: *17 August 2020*
Published: *23 October 2020*

Citation:
Chen X, Huang C and Cheng Y (2020) Identifiability, Risk, and Information Credibility in Discussions on Moral/Ethical Violation Topics on Chinese Social Networking Sites.
Front. Psychol. 11:535605.
doi: 10.3389/fpsyg.2020.535605

One heated argument in recent years concerns whether requiring real name supervision on social media will inhibit users' participation in discoursing online speech. The current study explores the impact of identification, perceived anonymity, perceived risk, and information credibility on participating in discussions on moral/ethical violation events on social network sites (SNS) in China. In this study, we constructed a model based on the literature and tested it on a sample of 218 frequent SNS users. The results demonstrate the influence of identification and perception of anonymity: although the relationship between the two factors is negative, both are conducive to participation in discussion on moral/ethical violation topics, and information credibility also has a positive impact. The results confirmed the significance of risk perception on comments posted about moral/ethical violation. Our results have reference value for identity management and internet governance. Policies regarding users' real names on the internet need to take into account the reliability of the identity authentication mechanism, as well as netizens' perceptions of privacy about their identity and the necessity of guaranteeing content and information reliability online. We also offer some suggestions for future research, with a special emphasis on applicability to different cultures, contexts, and social networking sites.

Keywords: anonymity perception, risk perception, information credibility, content moderation, real names on social media

INTRODUCTION

The complex integration of the internet and the real world means that in both the West and China, cyberspace has become the most convenient place for free expression, which is constrained by social norms and conformity (Lipschultz, 2018). Online public opinion is becoming the mainstream public opinion domain in China (Yu, 2017). China arguably presents an interesting case study on social networking sites (SNS) because it limits social media communication on non-domestic sites, establishing a microcosm of SNS (Sullivan, 2014). The expression of online public opinion is rooted in the social and cultural background of real-life society. In Chinese culture, there has always been an emphasis on "denying self and returning to propriety", personal behaviors should be "gentle, modest and courteous" and expressions should be humble and low-key (Chen, 2014).

In interpersonal communication, a superficial balance of relationship should be pursued, and telling the truth should be avoided to prevent harming interpersonal relationships (Zhai, 1999). In fact, culture is shaped by reality. When real lives are mapped onto virtual cyberspace in a hidden form, this principle of superficial balance is no longer important (Chen, 2018). Since Chinese people lack freedom of expression of their real views with their real-life social contacts, online anonymity is of greater importance to Chinese people compared with those from the West. In an interview survey conducted in 2017, 79.1% of 48 respondents said that they assume different identities online, which is reflected in using different SNS accounts (Chen, 2018).

In the most recent couple of decades, many researchers have regarded anonymity as directly enabling free expression on the internet as well as being the root cause of anomie (Nissenbaum, 1999; Davenport, 2002; Kim et al., 2011; Salanova et al., 2013; Stroud, 2014). The system of using real names online, which is considered a way of enhancing oversight of cyberspace and regulating the behavior of netizens, has been gradually established and improved with the development of China's internet governance (Lin, 2010; Liu, 2013).

In 2015, the State Internet Information Office of China issued a regulation named Rules on Account Name of Internet Users, which requires all users to submit real-identity registration information when using the internet. With the precondition that internet regulators can confirm users' identities, users have the right to use virtual names in online public speech spaces, which should be respected (Chen and Li, 2013). The real-name system is a mechanism that enables an individual's name to be mapped to that person's identity on social media. Users must provide information on their real personal identity when engaging in online activities, so as to establish a consistent relationship between their online and offline identities, enabling a confirmatory mechanism that links the rights, obligations, and interests of individuals' words and deeds online and in real life (Chen and Li, 2013).

China's enforcement of the online real-name registration system sparked widespread and fierce disputes, focused on its impact on netizens' freedom of expression. Supporters of the regulation argued that the system is conducive to creating a credible online speech environment and encouraging people to be responsible for their own speech. For those who are willing to speak frankly, real-name speech can also improve personal credibility and give weight to their words (Huang and Zhang, 2010). Opponents contend that the real-name system undermines the traditional values of equality, freedom, and openness on the internet, discourages internet users from participation in politics and scrutinizing government, and poses a covert threat to netizens' right of "freedom of expression" (Zhang and Lu, 2010).

Public concern about things that affect the majority of society is an important force in implementing oversight and promoting social progress. In China, Weibo and WeChat, with 500 million and 1 billion active users respectively at the beginning of 2020, have become the two most important SNS for people to express public opinion, offering different kinds of platforms for open and critical debate (Rauchfleisch and Schäfer, 2015).

In recent years, China has experienced many public opinion incidents online, with some incidents (both online and offline) sparking a great deal of online reaction and widespread discussion. The vast quantity of views freely expressed online by the public on specific topics has promoted social regulation offline, including efforts to promote the optimization of the social system and to combat corruption (Liebman, 2011). This plays an important role in social justice and promoting reform of the system of governance in the real world. The ability to trace a person's identity magnifies the risk of individual participation in exposing social problems, including interpersonal risks, moral risks, and even security risks, and this is an important reason for interrogating the online real-name system. However, to date studies on identifiability online have failed to explore this aspect. Therefore, the first research question to be tested in the current study is:

> *RQ1: Does the traceability of network identity information inhibit public participation in discussions on moral/ethical violation topics by internet users in China?*

Anonymity in cyberspace is an important way of protecting private information (Brazier et al., 2004; Rainie et al., 2013) and is conducive to the construction of self-image (Lin and Utz, 2017). The emergence of new cyber applications has led to a heated debate over the advantages and disadvantages of anonymity in cyberspace (Chen et al., 2016, 2019a,b; Christopherson, 2007; Lapidot-Lefler and Barak, 2012; Scott and Orlikowski, 2014; Fox et al., 2015; Jardine, 2015; Levontin and Yom-Tov, 2017). A series of research studies have confirmed that the perception of anonymity has different impacts on behavior in different online environments (Jessup et al., 1990; Joinson, 2001; Reinig and Mejias, 2004; Lapidot-Lefler and Barak, 2012; Yoon and Rolland, 2012; Hsieh and Luarn, 2014).

In the current cyberspace environment, absolute anonymity does not exist (Bodle, 2013). The issue of anonymity is often the focus of research on free expression (Akdeniz, 2002). SNS, especially those in which users tend to use real names, such as Facebook and WeChat, provide users with the freedom to make choices; the social connections built and maintained by these platforms may reduce the perception of anonymity. The positive impact derived from a perception of anonymity on positive self-disclosure has been analyzed in detail (Chen et al., 2016). Positive self-disclosure relates to the construction of self-image. However, participation in the discussion of topics that violate ethics is related to social responsibilities. Everyone has the responsibility to assume the ethical responsibilities of the media (Boeyink and Borden, 2010). For Chinese who value harmony of interpersonal relationship and the dignity, online anonymity has become a "veil", creating conditions whereby they can express their opinions freely. Anonymity in cyberspace is of great significance for Chinese netizens to express free speech about their true views. There is a lack of analysis in the current literature on the impact of perception of anonymity on users' participation in assuming public social responsibilities. Therefore, the current study attempts to answer the following research question based on the SNS environment in China:

> *RQ2: Does the perception of anonymity on the internet help drive netizens' public participation in discussions on moral/ethical violation topics in China?*

Based on the understanding of the issues described above, we synthesized the existing literature to build a theoretical model, as well as referring to the theories of the social identity model of deindividuation effects (SIDE) and Borden's communication ethical rules. Empirical research was conducted to test the hypothesis and theoretical model. The current study makes two principal contributions to the literature. First, it reveals that identification and perception of anonymity are opposite aspects of the influence mechanism of online participation in discussions on moral/ethical violation topics, and encouraging such participation needs to take into account underlying aspects of identification and the psychological perception of anonymity at the surface. Second, it confirms the role of risk perception and information credibility in participation in discussions on moral/ethical violation topics.

The remainder of the paper consists of the following sections: Section 2 reviews the theoretical foundation, and Section 3 proposes our research model and hypotheses. Section 4 explains the methodology, while Section 5 presents the results and related analysis. Section 6 discusses the results of the current study, together with the theoretical and practical limitations and potential avenues and implications for future research.

THEORETICAL FOUNDATION AND LITERATURE REVIEW

Cyber Anonymity and Online Public Opinion

Anonymity means a lack of identification of one's real identity (Marx, 1999). As a result of the integration between the internet and the real world, identification of users' real identity has become the basis for internet services and governance in China (Chen, 2018). According to the social identity model of deindividuation effects (SIDE), in the context of anonymous identities, people show a behavioral tendency to obey a group norm due to the prominence of an individual identity (Vilanova et al., 2017). The effects of online anonymity shown in the SIDE model are reflected in personalization, misconduct, and false information, which are related to the dark side of cyberspace (Fox and Moreland, 2015). However, anonymity in cyber-based communication may not necessarily lead to antisocial behavior (Christopherson, 2007). In some scenarios, anonymity enhances social processes related to group identity in online communication (Spears, 2017). On SNS, anonymity can also play a positive role in information exchange (Yoon and Rolland, 2012; Chen et al., 2016).

In the 1990s, use of cyber-based communication technology facilitated an anonymous communication environment, but this positive outcome is no longer the case (Chen et al., 2016). This is because anonymity is now seen by some as dangerous due to the following factors: issues in the protection of business transaction security (Chen et al., 2019a); government oversight and control; concerns about intellectual property; national and international legal implications; and the use of identity management technology (Froomkin, 2015). In SNS, user identification takes complicated forms, with complex and diversified functions and methods of interpersonal interactions.

The importance of online opinion, also called online word of mouth (e-WOM), has been confirmed (Goldsmith and Horowitz, 2006; Weeks et al., 2017). Despite the increasing disappearance of anonymity on the internet, it is still an important "safety valve" for the oppressed, dissidents, and whistleblowers to speak freely (Froomkin, 2015). This ability often comes from the psychological perception that they can engage in free speech without fear of the consequences (Chen et al., 2016). Therefore, online opinion makers can continue to participate, innovate, and explore topics and issues with a high degree of self-cognition. They boast stronger computer skills and use the internet more frequently (Lyons and Henderson, 2005).

Many studies have demonstrated a positive relationship between internet use, online speech, and political participation (Shah et al., 2005; Van den Eijnden et al., 2008; Valenzuela, 2013; Boulianne, 2015). In China, although government censorship inhibits people's willingness to voice their opinions to some extent, thanks to its loose network structure, which provide users with flexible expression forms and places to disclose opinions, the internet has still led to progressive changes in Chinese society (Shen et al., 2009). Public opinions on the internet affect the real world through users' discussion of specific events and dissemination of information (Yue et al., 2017). The main participants in cyberspace include stakeholders and the public (Zhang et al., 2015). The methods of participation include providing information, making comments, and involvement in decision-making or particular behaviors. The results of participation affect public decision-making or governance behaviors. Researchers have studied the impact of cyber anonymity on self-disclosure and information sharing (Yoon and Rolland, 2012; Chen et al., 2016, 2019b).

China is in a social transformation period, bringing a high degree of uncertainty to people's lives. The prevalent practice of concealing their true views for self-protection makes Chinese netizens present more complex mentalities and more diversified modes of behavior than before the construction of cyber identity (Chen, 2018). This also suggests that anonymous expression online plays a crucial role in alleviating potential pressure in real society and relieving the latent contradictions and conflicts. However, no empirical research has been conducted on the impact of a lack of anonymity on expression of public opinion on moral/ethical matters. The exploration of this impact mechanism is an important basis for establishing identity management and carrying out governance in cyberspace.

Participation in Moral/Ethical Oversight

The development of global media brings urgency to intercultural communications on ethics-related topics (Borden, 2016). Cyberspace transfers the function of traditional media and its societal influence from professional journalists to every netizen. Every speaker in cyberspace has the function of the media to some extent. Therefore, the discussion of ethical and moral responsibility in journalism theory provides an important reference point for individuals' posting information and sharing behavior on SNS. From a moral/ethical standpoint, journalists have to be clearly aware of what they are and what they are not, and whether they are to stand in favor of some things and against others (Borden, 2007). Media ethicist Elliott (1986) suggests that three levels of responsibilities provide the foundations for moral excellence in journalism: general responsibilities, particular responsibilities, and individuals' personal responsibilities.

Media participants should follow three ethical rules: truth telling, privacy, and fairness (Boeyink and Borden, 2010). In some controversial ethical violations, the three principles may come into conflict (Boeyink and Borden, 2010). However, in some cases involving ethical and moral principles in which people reach a consensus, the behavior of paying attention to and getting involved in the discussion itself is in compliance with the above three principles. Moral excellence consists of performing your ethical responsibilities well: all of us have moral responsibilities, such as to be truthful to avoid harming others and to keep our promises, so called general responsibilities, which matter in everyone's lives. These responsibilities should give individuals the power to supervise and condemn those behaviors that violate morality and ethics and endanger the foundation of human existence.

In cyberspace, netizens participate in discussion of an event to supervise and condemn behavior which violates norms of ethics and morality (Repnikova, 2017). The power of moral supervision plays an important role in aspects such as maintaining social justice, promoting improvements in the social system, and restraining corruption (Liebman, 2011). Even though our individual actions are constrained by general and particular responsibilities, media participants have to retain autonomy as moral agents (Elliott, 1986). In particular, when events occur that violate the universal morality of humanity, the involvement of people in discussion and information sharing in cyberspace strengthens the argument for justice; thus, such participation plays an important role in maintaining universal ethics and the morality of the social system. The current study also discusses how the concealment and revelation of individuals' real identity in cyberspace affect users' involvement in moral supervision.

Perceived Risk and Information Credibility Sources

Studies of risk perception examine the judgments that people make when they are asked to characterize and evaluate hazardous activities and technologies (Slovic, 1987). Perceived risk has been conceptualized in terms of the expected negative utility of particular actions (Peter and Ryan, 1976). The impact of perceived risk on behavior has been confirmed in online research contexts, including information sharing and control (Gerlach et al., 2015; Hajli and Lin, 2016) and adoption behaviors (Featherman and Pavlou, 2003; Horst et al., 2007; Martins et al., 2014; van Winsen et al., 2016). Perceived risk reduces users' perceptions of value (Snoj et al., 2004; Chang and Tseng, 2013) and destroys trust (Slovic, 1993). However, the influence of perceived risk on the expression of ethical views has been neglected in current research.

Perceived risk has an impact on people's moral judgments Subjects in a high-risk treatment group exhibited significantly harsher ethical judgments than those in a low-risk treatment group (Cherry and Fraedrich, 2002). Reputation, a sense of belonging, and satisfaction from helping others are significantly related to e-WOM intention (Cheung and Lee, 2012). Paying attention to moral views and participating in social media posts and sharing this kind of information is of concern for the collective interest (Earle and Siegrist, 2006). The discussion of moral issues online is related to the altruistic punishment mechanism of human behavior, which holds that individuals voluntarily take risks and pay costs for punishing people who violate social norms, and this plays an important role in the evolution of social cooperation (Fehr and Fischbacher, 2004; Boyd et al., 2010).

The framing of risk depends on the media used to perceive it (Ericson and Doyle, 2003). In seeking information, people rely on information sources to build trust, which is in play whenever users exchange information, and the information source—the trusted party—may have a moral responsibility to an information seeker (Hertzum et al., 2002). Information credibility has become an important topic as the internet has become increasingly ubiquitous (Kelton et al., 2008). The influence of trust in digital information has been confirmed as a key mediating variable between information quality and information use (Pan and Chiou, 2011), but the influence of information credibility on information related to discussions on moral/ethical violation topics needs further clarification.

RESEARCH MODEL AND HYPOTHESIS

The current study constructs a model to examine how the factors of identity perception, real names, and perceived anonymity affect the intention to participate in SNS discussions on moral/ethical issues in China. In addition, the current study also explores the influence of risk perception and information credibility on participation in social media discussions on moral issues in China. The research model is depicted in **Figure 1**.

Identification and Anonymity on Social Networking Sites

User identifiability and perception of anonymity are not two sides of the same factor. Although they have opposing properties, they are two different factors. The ability to identify internet users refers to the process and the potential for identifying the true identity of users in cyberspace; this is not just a legal concept but also a technical means of identity detection with the help of publicly available methods (Krausová, 2009). In cyberspace, user identifiability is represented by the richness of information in terms of whether there are clues to determine a user's real identity (Chen et al., 2019a). Identifiability is objective, whereas perceptual anonymity is subjective and is the psychological

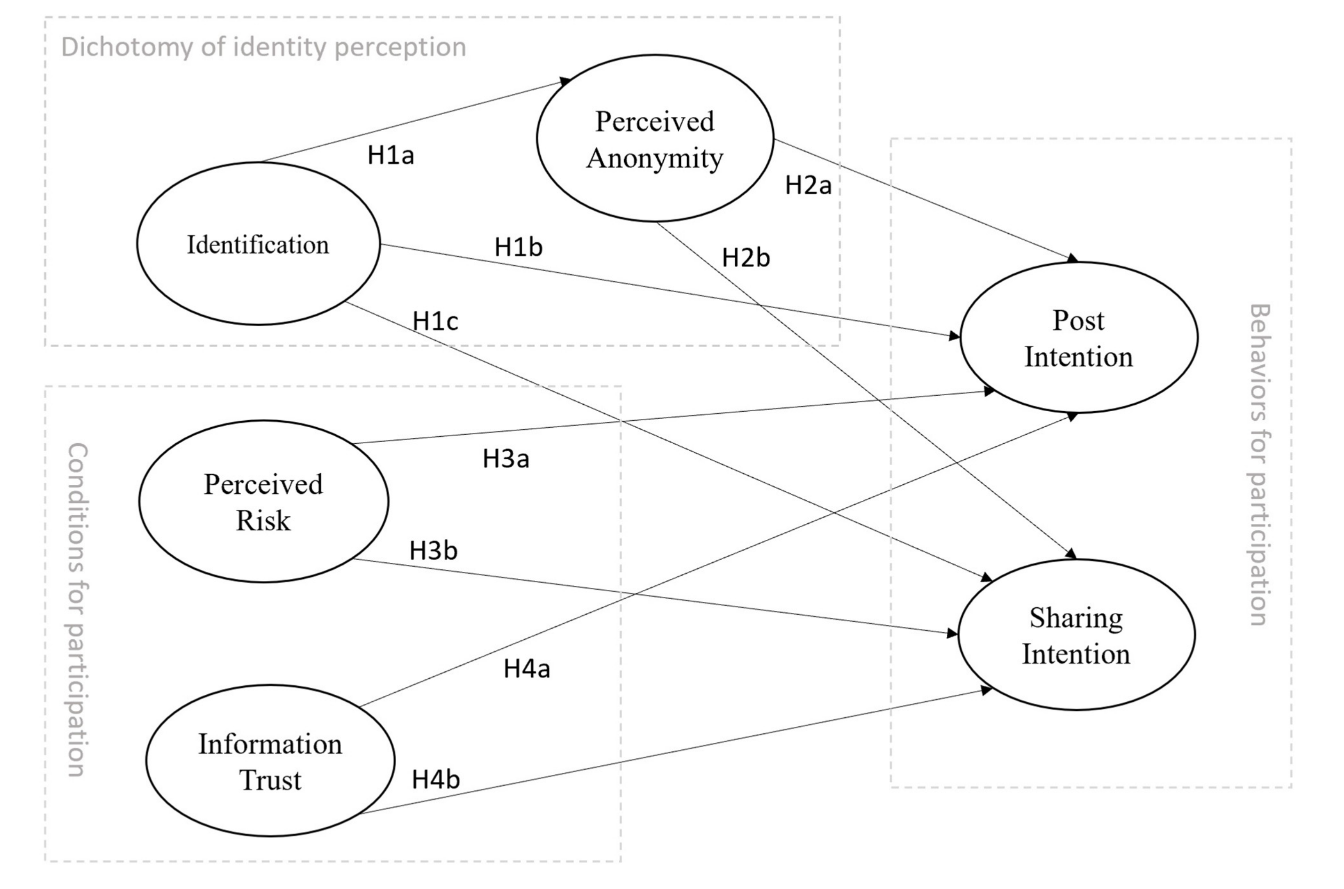


FIGURE 1 | Research model.

perception of the nature of the subject. Previous research has confirmed that identity has a significantly negative impact on perceptual anonymity on various online social media platforms (Chen et al., 2016, 2019a,b). The relationship between these two factors should be the same in monitoring and participating in discussions on moral/ethical violation topics on online social media. Thus, we hypothesize that:

H1a: User identification has a negative impact on their perception of anonymity in discussing moral/ethical violation topics on Chinese SNS.

Identity is perceived as a social process that aligns with internal self-identification and external identity classification (Jenkins, 2014). Identification and self-efficacy are closely intertwined, and the connection between social identity and self-efficacy is further supported by social identity theory (Guan and So, 2016). Four factors—performance accomplishments, vicarious experience, verbal persuasion, and emotional arousal—contribute to a boost in self-efficacy (Bandura, 1977). Social influence and perceived control will positively impact self-disclosure in SNS (Cheung et al., 2015). External factors, such as environment and information input, appear to affect self-efficacy through their influence on internal variables, such as motivation, ability, or performance strategies (Gist and Mitchell, 1992). The identity construction of the user depends on the creation and sharing of information. Attention to and discussion of online public opinion are also a way for the user to construct his or her own identity. The more active the user is on online social media, the deeper the recognition of the user's online identity and thus the greater enthusiasm and self-efficacy the user has for participating in discussion on the topic. Thus, we hypothesize that:

H1b: User identification has a positive impact on comment intention in discussing moral/ethical violation topics on Chinese SNS.

H1c: User identification has a positive impact on sharing intention in discussing moral/ethical violation topics on Chinese SNS.

Perception of anonymity refers to the indiscernibility of the identity of the user, which leads to self-awareness of identity anonymity, that is, that one cannot be tracked in cyberspace (Kang et al., 2013). In an anonymous environment, social bonds are weaker, and social norms tend to be enforced more aggressively (Wright, 2014). The relationship between perception of online anonymity and behavior depends on the specific communication context (Joinson, 2007). In group discussions, it has been found that users who perceived anonymity were more likely than identified users to embellish the opinions of others (Jessup et al., 1990). For users, circumventing the possibility of authentication can protect privacy (Brennan et al., 2012). Although online anonymity introduces uncertainty into interpersonal interactions, it also reduces risks in online privacy and security (Rainie et al., 2013). According to the SIDE theory, within an anonymous context, people tend to comply with collective norms. Following the argument that general moral/ethical principles lead to collective behavior and consensus (Boeyink and Borden, 2010), the SIDE effect will promote the individual's obedience to collective behaviors. In discussion and online decision-making on certain sensitive topics, online anonymity increases behavioral contributions and effective suggestions (Jessup et al., 1990). In addition, research on perception of anonymity in the use of online social media for information sharing found that perception of anonymity has a positive effect on self-disclosure (Chen et al., 2016). Thus, we hypothesize that:

H2a: Users' perceived anonymity has a positive impact on comment intention in discussing moral/ethical violation topics on Chinese SNS.

H2b: Users' perceived anonymity has a positive impact on sharing intention in discussing moral/ethical violation topics on Chinese SNS.

Perceived Risks Online

Perceived risk has been conceptualized in terms of the expected negative utility of actions (Peter and Ryan, 1976). Risk discourse is redolent with the ideologies of mortality, danger, and divine retribution (Lupton, 1993). Participation in familiar activities has a tendency to minimize the probability of bad outcomes (Douglas, 2013). Decisions about risk as moral decisions are made in the context of uncertainty (Adams, 2003). However, risk perception has different influence mechanisms in play in discussions on moral/ethical violation topics. The increase in risk entails an attendant enhancement of new moral responsibilities at multiple levels in a society (Ericson and Doyle, 2003). From a moral/ethical standpoint, the media participant has to be clearly aware of their responsibilities (Borden, 2007); in general moral/ethical events in particular, especially the event challenging the basic value and living of human being. the basic principles should be clear (Elliott, 1986). Three ethical rules—truth telling, privacy, and fairness—may come into conflict (Boeyink and Borden, 2010), and should be taken in consideration. Even though our individual actions are constrained by general and particular responsibilities, media participants have to retain autonomy as moral agents (Elliott, 1986). We believe that perception of risk in cyberspace should have a positive impact on participation in discussions on moral/ethical violation topics. Thus, we hypothesize that:

H3a: Users' perceived risk has a positive impact on comment intention in discussing moral/ethical violation topics on Chinese SNS.

H3b: Users' perceived risk has a positive impact on sharing intention in discussing moral/ethical violation topics on Chinese SNS.

Information Credibility

Trust in technology is constructed in the same way as trust in people (McKnight, 2005). Information credibility is a descriptive factor of perceived information quality which influences information exchange. Information quality during an exchange can help build trust and reduce perceived exchange risk (Nicolaou and McKnight, 2006). External factors, such as environment and information input, appear to affect self-efficacy through their influence on internal variables (Gist and Mitchell, 1992). Self-efficacy factors, such as perceived performance, have been confirmed as having an adverse impact on the adoption of e-services (Featherman and Pavlou, 2003), and are also positively related to a consumer's trust expectation (Hong, 2015). The perceived trustworthiness of information determines the level of confidence developed by the user and the corresponding willingness to use the information (Kelton et al., 2008). Research has also determined the importance of trust in forecasted information sharing in television, newspapers, and online news (Kiousis, 2001), and supply chains (Özer et al., 2011). The source of information is an important factor in considering information credibility (Lucassen and Schraagen, 2010). Morality—relevant information provides the check for value similarity and generates trust (Earle and Siegrist, 2006). The impact of trust on forecasted information sharing has been confirmed (Zimmer et al., 2010; Ha and Ahn, 2011; Özer et al., 2011). Thus, we hypothesize that:

H4a: Users' information credibility has a positive impact on their comment intention in discussing moral/ethical violation topics on Chinese SNS.

H4b: Users' information credibility has a positive impact on sharing intention in discussing moral/ethical violation topics on Chinese SNS.

METHODOLOGY

The internet is an ideal medium for collecting data from different groups (Koch and Emrey, 2001). The current study focuses on Chinese SNS users' participation in discussions on moral/ethical topics. In China, WeChat and Weibo are the most popular online social platforms that people use to express their opinions (Hou et al., 2018). According to the Social Global Web Index's flagship 2018 report on the latest trends in social media, Facebook is the world's largest SNS with more than 2.6 billion users, WeChat is ranked fourteenth, and Weibo is ranked eighteenth; the latter are also the only two Chinese social media platforms listed in the global top 20 (Global Web Index, 2018). At the end of 2019, WeChat had over 1 billion active users around the world. Weibo, which has more than 500 million active users, provides a virtual public space for users to share their opinions with their connected peers, making it the most influential opinion platform in China; this makes it a suitable arena for our research on participation in discussions on moral/ethical topics. To validate our hypotheses, we conducted a survey on the use of WeChat and Weibo by Chinese users.

The current research is based on a heated event which provoked discussion all over the world, in which a scientist named He Jiankui announced his work on editing the genes of a fetus. Scientists and authoritative academic institutions from different countries gave their opinions, arguing that He Jiankui had seriously violated academic morals and the code of conduct. What he did also caused an outcry in the international community (Cyranoski and Ledford, 2018; Normile, 2018). His behavior did not only violate scientific ethics, but also had the possibility of polluting the human gene pool and posed a threat to the future of humanity. It was noted that the ethical infractions in this work are among the most egregious that have been recorded in modern medical history since the Second World War (Kuersten and Wexler, 2019). This ethical and moral violation event is significant for the whole of humanity.

Data were collected for the current study at the point in time when this gene-editing of a fetus had just occurred, which had attracted the attention of the whole world and had become a heated topic on various SNS. This was important for focusing the participants' attention on the research issues, and to obtain a clearer understanding on the event. The administered questionnaire consists of three parts, the first of which is a privacy and protection statement and informed consent declaration. The participants read the information carefully and confirmed it. The second part consists of two news reports about the gene-editing fetus event from *People's Daily*, which is the most authoritative official media source in China, and the *Beijing News*, which has a wide influence and is based in Beijing. The reports (a total of 884 words) described the development of the event up to December 18, 2018, providing the objective facts calmly and without emotional appeals. The third part required the participants to complete a survey about the gene-editing situation and their feelings about it.

Data Collection

At the preliminary stage, the current study tested the research model through an investigation of frequent SNS users in China. The aim was to answer the research questions about how users' identity impacts participation in online speech on moral/ethical violation topics on SNS. According to the relevant institutional and national guidelines and regulations, ethics approval was not required. First, the data collection of the current study did not involve implication, drugs, or mental manipulation, as the participants were only required to report their experiences and behavioral tendency according to their SNS use conditions. Thus, no issues with respect to safety, health, or protection of rights and interests were involved. Second, the data collection required no identity concealment, as no privacy or sensitive issues were involved. Third, the questionnaire for collecting the data in the current study includes an informed consent statement, and the participants were only requested to answer questions according to their SNS use conditions. The data were only to be used for scientific research, without influencing the privacy, reputation, living conditions, or health of the participants. Fourth, the potential participants were offered anonymity; they were fully aware of this option before, during, as well as after giving their responses. Fifth, the current study did not store or use the private information of participants, and any information that may lead to identity risks (only the IP address) was removed during the analysis and submission for scientific review.

The data were collected with a questionnaire using a sample service provided by an online survey platform (wjx.cn/sample/service.aspx). This is the largest online survey agency in China, providing 2.6 million sample banks consistent with the demographic distribution of China's netizens. The survey employed a purposive sampling method focused on the frequent users of Chinese SNS. Based on user requirements, the platform sends the invitation email to randomly selected potential participants from the sample banks. The survey would begin once the potential participants clicked the URL in the email. First, they were asked to read a news article about the topic, and then they were asked to fill in a questionnaire online. To identify frequent users of social networking platforms for the study and to confirm the quality of data, we included screening questions and limited the response time as a filter, which yielded a total of 218 valid questionnaires out of 345 responses. The service provider was paid 1 USD for each valid sample. We used the SmartPLS to conduct empirical research. One of advantage of using SmartPLS is the sample size,Some SEM based methods need samples of at least 200 samples or more,but SmartPLS is suitable for sample sizes of less than 200 (Sander and Phoey, 2014). **Table 1** shows the demographic characteristics of the respondents to the survey.

TABLE 1 | Demographic characteristics of respondents.

Characteristic		Number (%)
Gender	Female	149 (68.3)
	Male	69 (31.7)
Age (y)	Younger	0 (0)
	15–24	56 (25.7)
	25–34	109 (50.0)
	35–44	38 (17.4)
	45–54	12 (5.5)
	Older	3 (1.4)
User history	< 1 year	0 (0)
	1–3 years	3 (1.4)
	3–6 years	45 (20.6)
	6–10 years	90 (41.3)
	> 10 years	80(36.7)
Education level	Primary school and below	0 (0)
	Junior middle school	1 (0.5)
	Senior middle School	5 (2.3)
	Technical Secondary school	6 (2.8)
	Junior college	37 (16.9)
	Bachelor's degree	147 (67.4)
	Master's degree	21 (9.6)
	Ph.D.	1 (0.5)
	Other	0 (0)
Marital status	Unmarried	92 (42.2)
	Married	125 (57.3)
	Divorced	1 (0.5)

The sample had a reasonable demographic distribution. Referring to the data in reports published by affiliated companies Sina and TenCent on Weibo and WeChat in 2018, the age distribution characteristics of the sample were basically consistent with those of Weibo and WeChat users. In the sample, respondents with a bachelor's degree or above accounted for 77.5% of the total; the education level was slightly higher but within a similar range to that of Sina Weibo users in the reports, 70.8% of whom had university degrees. In WeChat (64% male users) and Weibo (57% male users), the proportion of male users was higher than that of female users. Although the gender distribution of samples may generally lead to bias in the results, gender differences do not affect research on general online sharing behaviors, as confirmed by some related studies on gender distribution differences (Yoon and Rolland, 2012; Cheung et al., 2015; Yan et al., 2016). In addition, data from the current study show that SNS users prefer to browse information rather than publish information. The participants spend on average 81.2% of their time browsing information and 18.8% of their time publishing information and participating in discussions.

Measures

All measures were adapted from well-established scales, the validity of which had been confirmed in the relevant existing literature. Multi-item measures were applied to ensure the validity and reliability of the study. To ensure comprehension by Chinese users, we translated the scale into Chinese and then back-translated it into English. We asked two researchers to verify the consistency of the terms used in the scale to ensure that the translation and terms were consistent. The scale was modified slightly to fit the SNS context. A seven-point Likert scale (1 = strongly disagree to 7 = strongly agree) is used in all measures. **Table 2** lists the constructs and measures applied in the research, as well as the source references. The psychometric properties in **Table 2** include Cronbach's α, composite reliability (CR), and average variance extracted (AVE) of the constructs, as well as the loading, *T*-value, mean, and standard deviation (SD) of the measure items used in the current study.

RESULTS

We used a structural equation model to verify the research model and performed statistical analysis using the partial least squares method (PLS). In addition, we used Smart PLS version 3 (Ringle et al., 2014) to test the research model empirically; this is an analytical technique widely used in social science research because it provides a flexible and exploratory method with coherent explanations of complex relationships (Henseler et al., 2014). In accordance with the two-step analysis method (Hair et al., 2006), we tested the credibility and validity of the measured values and then evaluated the structural model.

In the next sections, we analyzed the data in two steps: first, the measurement and data were tested for reliability and validity, and then we drew conclusions about the structural relationship based on the measurement instruments with desirable psychometric properties.

Reliability and Validity of the Measurement Items

As shown in **Table 2**, all the indicator loadings were significant and higher than 0.70, except ID1 and PA1, whose loadings were lower than 0.7; therefore, we dropped them, ensuring the convergent validity of the measurement model. The resulting Cronbach's α of each construct exceeds the recommended level (0.70), and the composite reliability is higher than 0.80, indicating that the reliability of all latent variables is very good. In addition, each variable has good polymerization validity, because the AVE of all latent variables surpasses 0.6.

Ensuring discriminant validity requires a low correlation between measures and other structural measures (Fornell and Larcker, 1981). In **Table 3**, the main diagonal value is the square root of AVE and the out-of-diagonal value is the correlation coefficient between the constructs. All the diagonal values are higher than 0.7 and exceed the correlation between any pair of measures. This value indicates that the model also has good discriminant validity. Therefore, the results of our data analysis have adequately high discriminant validity.

Structural Model

Before testing the hypotheses, multicollinearity regarding the structure of the data was tested and was in accordance with the requirements. We then examined the structural model by analyzing the significance of the path coefficients and the R2 variance for the dependent constructs based on the hypothetical research model. The path and its importance for the structural model, the coefficients of each related structure, as well as their *T*-values on the structural model and the deterministic coefficients (R2) are illustrated in **Figure 2**.

For the full model, most proposed hypotheses are strongly supported by empirical evidence with significance at $p < 0.05$, except for H3b. In this section, we discuss how each construct in our theoretical framework influences the two types of participatory behaviors. Regarding identifiability, we found that identification has a strongly negative influence on perceived anonymity ($\beta = -0.566$, $p < 0.001$). This finding is consistent with results in the previous literature; those people who are objectively identifiable will not perceive themselves to be anonymous. Thus, H1a is supported. The results also show that identification leads to participatory behaviors in discussion on moral/ethic violation topics, both in terms of comment intention ($\beta = 0.321$, $p < 0.01$) and sharing intention ($\beta = 0.365$, $p < 0.001$). In the SNS context, the more personal the information is that is disclosed, the more likely the user is to attend to the discussion and carry out moral/ethical supervision. Therefore, H1b and H1c are supported. Perceived anonymity also leads to participation in discussion on related topics, both in terms of comment intention ($\beta = 0.261$, $p < 0.01$) and sharing intention ($\beta = 0.205$, $p < 0.05$). Perceptions of anonymity also declaim importance in participation in related activities. Therefore, H2a and H2b are supported. Furthermore, the coefficient of the path

TABLE 2 | The measures and psychometric properties.

Items	Loading	*T*-value	Mean	SD
Identification (Chen et al., 2019a) (Cronbach α = 0.828; CR = 0.886; AVE = 0.661)				
ID1: I revealed my real name on my social media account. (dropped)				
ID2: I may reveal my name in the messages I post on the social media account.	0.761	20.125	4.252	1.616
ID3: You'll probably know who I am from my social media accounts.	0.858	43.521	4.344	1.602
ID4: The content I post on my social media accounts is very personal, and it's easy to tell who I am.	0.867	49.199	3.940	1.548
ID5: I revealed some social information about myself in my social network account, such as company, age, occupation, and hobbies.	0.761	22.251	4.477	1.609
Perceived Risk (Chen et al., 2016) (Cronbach α = 0.873; CR = 0.897; AVE = 0.638)				
PR1: I am concerned that participating in this discussion will adversely affect my personal fortunes.	0.738	5.243	3.784	1.386
PR2: I am concerned that participating in this discussion will adversely affect my use of this account.	0.717	4.259	4.032	1.516
PR3: I am concerned that participating in this discussion will adversely affect my personal safety.	0.815	5.982	3.642	1.779
PR4: I am concerned that participating in this discussion will adversely affect my mental state.	0.847	7.708	3.408	1.646
PR5: I am concerned that participating in this discussion will lead to backlash from my family, friends, and acquaintances.	0.864	7.092	3.500	1.654
Information Credibility (Li and Suh, 2015) (Cronbach α = 0.875; CR = 0.914; AVE = 0.726)				
IC1: I think the source of this incident is believable.	0.837	29.293	4.752	1.178
IC2: I think the source of this information is usually factual.	0.885	51.474	4.789	1.296
IC3: I think the source of the information about this incident came from credible sources.	0.819	15.792	4.601	1.365
IC4: I think the source of this information is trustworthy.	0.867	44.161	4.638	1.359
Perceived Anonymity (Hite et al., 2014) (Cronbach α = 0.883; CR = 0.919; AVE = 0.739)				
PA1: I believe that people who can see what I post or share don't know who I am. (dropped)				
PA2: I think that people who can see what I post or share don't know who I am.	0.839	30.401	3.688	1.466
PA3: It is likely that my account will reveal who I am. *	0.870	53.487	3.417	1.510
PA4: Some one else who could see my posting would know my true name. *	0.880	49.815	3.422	1.587
PA5: My personal identity can be guessed by others. *	0.850	34.114	3.252	1.531
Comment Intention (Jang et al., 2016; Kwon, 2020) (Cronbach α = 0.784; CR = 0.860; AVE = 0.606)				
bCI1: I will try to post comments on this event.	0.768	23.405	4.193	1.408
CI2: I tend to comment on my friends' post.	0.828	34.585	4.193	1.440
CI3: I intend to comment on the event more frequently.	0.763	18.592	4.587	1.428
CI4: I will always make an effort to comment on it.	0.754	16.273	3.954	1.667
Sharing Intention (Chung et al., 2016) (Cronbach α = 0.892; CR = 0.925; AVE = 0.756)				
SI1: I am inclined to forward reports on the incident to others on my SNS.	0.860	35.704	4.413	1.525
SI2: I tend to post this event to let others on my SNS know about it.	0.845	35.737	4.560	1.517
SI3: I will share this event to let others on my SNS know about it.	0.906	73.530	4.101	1.686
SI4: I usually spread news about this event to others on my SNS.	0.865	42.843	3.982	1.684

**Reverse scale.*

TABLE 3 | Correlation matrix and psychometric properties of key constructs.

	ID	PR	IC	PA	CI	SI
Identification (ID)	(0.813)					
Perceived Risk (PR)	0.064	(0.799)				
Information Credibility (IC)	0.263	−0.069	(0.852)			
Perceived Anonymity (PA)	−0.587	−0.023	−0.069	(0.860)		
Comment Intention (CI)	0.242	0.205	0.365	0.053	(0.779)	
Sharing Intention (SI)	0.301	0.137	0.337	−0.022	0.622	(0.869)

SQRT (AVE) is in parentheses. Off-diagonal cells show the correlations between constructs.

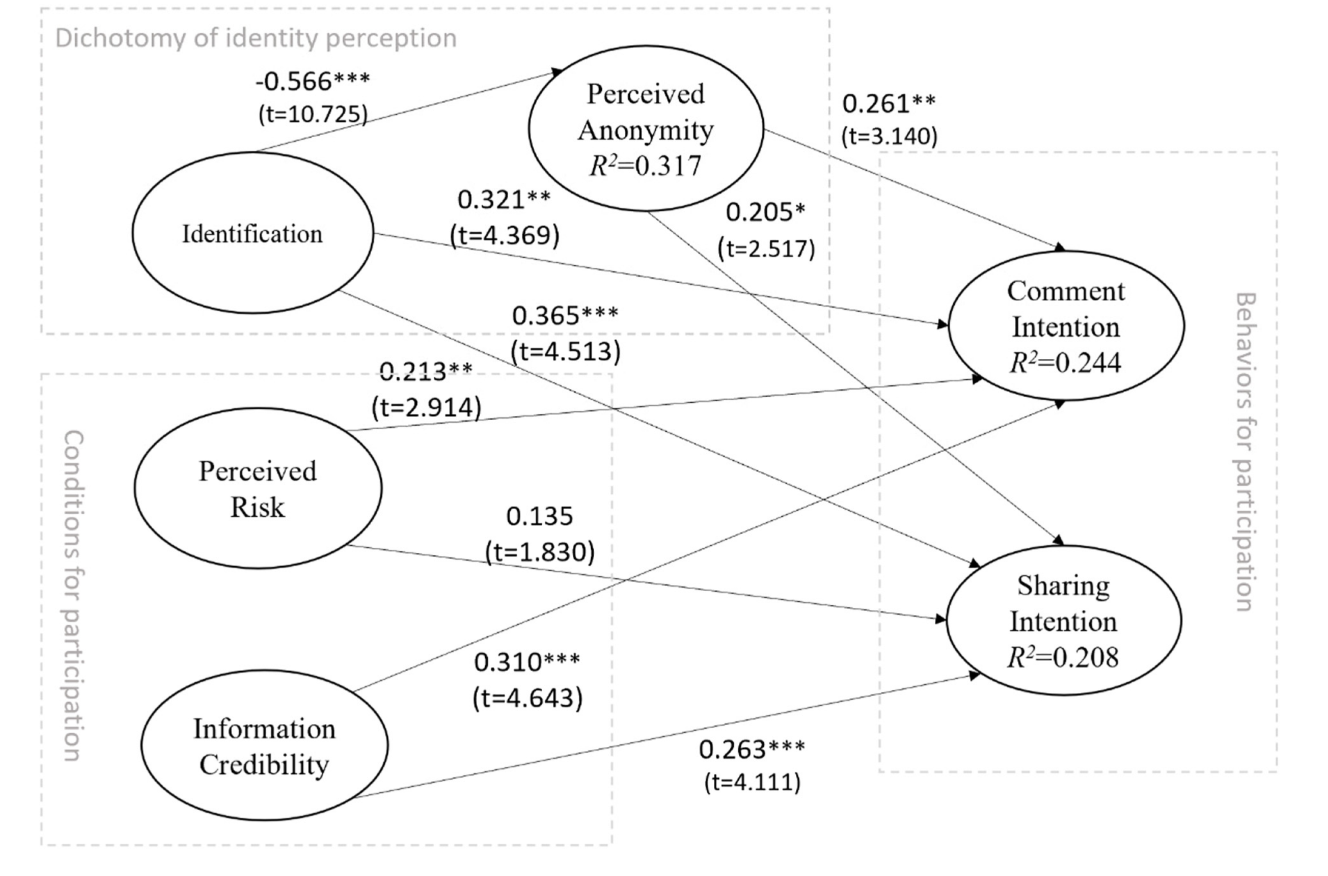


FIGURE 2 | Results of the research model.

from identification to sharing intention is greater than that from identification to comment intention. However, the coefficient of the path from perceived anonymity to comment intention is greater than that from identification to sharing intention. Sharing related information is helpful for the user's image, while making comments carries less interpersonal pressure. This finding suggests that identification and perceived anonymity both have a positive impact on participation in discussion on moral/ethic violation topics, but through a different influence mechanism.

Regarding the conditions for participation, perceived risk has a significant impact on comment intention (β = 0.213, $p < 0.01$), which is in line with the altruistic punishment mechanism of human behavior. Thus, H3a is supported. However, perceived risk has an impact on sharing intention with a significance level of $p = 0.067$, suggesting that H3b is not supported in the current study, while the direction of the impact is consistent with H3b. We can say that at the test standard of $p < 0.05$, the empirical research cannot significantly support H3b. Regarding perceived risk, people are more willing to comment on ethical and moral violations than to share information. From actual experience, the degree of exposure to information sharing in social networks is higher than that of commenting on information released by others. Without any doubt, information credibility leads to comment intention (β = 0.310, $p < 0.001$) as well as sharing intention (β = 0.263, $p < 0.001$). Trust in source information leads to confidence in a user's participation in discussion on

TABLE 4 | Direct, indirect and total effect (Bootstrap = 2000).

Effect Types		Effect Mean	S.E.	*T*-value	*P*-value
Total Effect	ID→PA	−0.566	0.053	10.725	0.000
	ID→CI	0.173	0.075	2.299	0.022
	ID→SI	0.249	0.074	3.364	0.001
Direct Effect	ID→PA	−0.566	0.053	10.725	0.000
	ID→CI	0.321	0.073	4.369	0.000
	ID→SI	0.365	0.081	4.513	0.000
	PR→CI	0.213	0.073	2.914	0.004
	PR→SI	0.135	0.074	1.830	0.067
	IC→CI	0.310	0.067	4.643	0.000
	IC→SI	0.263	0.064	4.111	0.000
	PA→CI	0.261	0.083	3.140	0.002
	PA→SI	0.205	0.082	2.517	0.012
Total indirect Effect	ID→PA→CI	−0.148	0.053	2.806	0.005
	ID→PA→SI	−0.116	0.050	2.307	0.021

related topics. The results imply the importance of information credibility even in moral/ethical violation topics. Therefore, H4a and H4b are supported.

The independent variables explain a substantial portion of the variance in the dependent variables. In the current model, identification explains 31.7% (r^2 = 0.317) of the variance in perceived anonymity, 24.4% (r^2 = 0.244) of the variance in comment intention with a significant impact from identification, perceived risk, information credibility, and perceived anonymity, and 20.8% (r^2 = 0.208) of the variance in sharing intention with a significant impact from identification, information credibility, and perceived anonymity.

In Smart-plus, Standardized Root Mean Square Residual (SRMR) and the Normed Fit Index (NFI) may assess the model fit. For SRMR, the recommended value should be lower than 0.08; NFI values between 0 and 1 are recommended. For the current model, SRMR is 0.067 and NFI is 0.781. The goodness of fit value of the model is 0.577, which is significantly higher than the standard of substantial fitting, in which 0.36, 0.25, or 0.1 can be described as, respectively, substantial, moderate, and weak (Marsh et al., 2005). The indices indicate an acceptable model fit of the data.

The results of direct effect (DE), total effect of each construct, and the results of indirect effects existing in the model, as well as the standard error and *T*-values of each effect are given in **Table 4**. The results show that all direct effects, except for the non-significant direct effect of perceived risk on sharing intention and the significant negative effect of identification on perceived anonymity (DE = −0.566), are positive and significant to varying degrees. Identification has the largest direct impact on comment intentions (DE = 0.321), followed by information credibility (DE = 0.310) and perceived anonymity (DE = 0.261), and perceived risk has the least direct impact on comment intention (DE = 0.213), but is still significant at $p < 0.01$. Among the direct influences on sharing intention, identification has the largest influence (DE = 0.365), followed by information credibility (DE = 0.263), and perceived anonymity has the smallest influence (DE = 0.205), but is still significant at $p < 0.05$.

Two significant total indirect effects have been identified in the model. If the sign of indirect effect is opposite to that of direct effect, the total effect will be suppressed (Wen and Ye, 2014). The suppressing effect of perceived anonymity accounts for 46.1% of the direct effect between identification and comment intention, and for 31.8% of the direct effect of identification and sharing intention. Perceived anonymity has a significant suppressing effect between identification and two participation factors.

DISCUSSION AND CONCLUSION

In the current study, we examined user participation in discussions on moral/ethical topics on Chinese SNS. To do so, we constructed a model to describe the influence of identification, perceived anonymity, perceived risk, and information credibility. The measurement model has been confirmed, with acceptable convergent and discriminant validity, path coefficients, and model fit.

Discussion of Results

Identifiability and perceived anonymity of SNS user identity are not two sides of an organic whole but, rather, two different elements (Chen et al., 2019a). Identifiability reflects the amount of information available on the real identity of the behavioral subject that is identified (Marx, 1999). High identifiability of users leads to low perceived anonymity of identity (Chen et al., 2016, 2019a,b), which is also suggested in the current study with the supportive result for H1a. Furthermore, with the supportive results for H1b, H1c, H2a, and H2b, the current study shows that the influence of online identification and perception of anonymity are both conducive to participation in discussion about moral/ethical violation topics. This result is in accordance with the research on self-disclosure on Weibo (Chen et al., 2016).

When users have control over their identification, perceived anonymity contributes to user participation in discussion on moral/ethical violation topics; when users have control over their perceived anonymity, identification also contributes to participation in discussion of moral/ethical issues. High user identifiability is advantageous in building a sense of identification with the social network identity and enhancing the credibility of the opinions expressed at the same time. The influence of perceived anonymity on speech behavior varies in different application scenarios; there is evidence of a negative influence on the perceived autonomy of sharing behaviors in cyberspace (Yoon and Rolland, 2012), whereas on social media it promotes self-expression without causing a reduction in the perception of self-expression risks (Chen et al., 2016). This leads us to conclude that both identification and perceived anonymity play important roles in participation in discussions of moral/ethical violation topics.

The current study also shows the positive effect of perceived risk on users' intention to post comments on moral/ethical topics on SNS in China: in the face of an event that raises common ethical concerns, when the level of risk perceived by users is higher, so too is their intention to post comments, as shown by the supportive result for H3a. Reducing cybersecurity risk increasingly depends on information sharing (Goodwin et al., 2015). This result is in line with the statement that risk may add value to SNS in some contexts, as users are motivated to reduce uncertainty (Mitchell, 1999). However, it is not consistent with some research studies which examined the impact of risk on information sharing behavior in other SNS contexts and showed no significant impact on self-disclosure (Chen et al., 2016), and that perceived privacy risk will negatively impact the attitude towards information sharing (Hajli and Lin, 2016). Despite the risks, people participate in relevant social activities to safeguard justice because of their moral sentiments (Gintis et al., 2005). This phenomenon reveals the particularity of participation in moral/ethical-related issues. Participating in discussions on moral/ethical violation topics is out of concern for fairness and justice, as well as to reduce uncertainty. This may be attributed to the neural basis of altruistic punishment in people's brains (Fehr and Gächter, 2002). Participation in discussions on moral/ethical violation topics on networked social media can be understood as reciprocal behavior with a price, because it is a kind of trial-and-punishment of behavior considered unethical, rather than behavior that is responsive or well-targeted. In the face of unethical events, the behavioral mechanism comes from people's desire to impose punishment and to gain a sense of satisfaction from participation in imposing punishment (De Quervain et al., 2004). Therefore, perceived risk does not make people shrink from discussion participation, but encourages them to participate in the discussion of moral and ethical issues to some extent. The encouragement from perceived risk does not have a significant effect on sharing intention, but the direction of the impact is consistent with H3b. The relations between perceived risk and intensive participation are of value to explore further.

We also found that information credibility significantly affects participation in discussions on moral/ethics violation topics. This has a positive influence on both posting comments and sharing intention on moral/ethics violation topics, with supportive evidence for H4a and H4b. This result is in accordance with the outcomes from research studies on the impact of information credibility on involvement in discussion and sharing (Zimmer et al., 2010; Ha and Ahn, 2011; Özer et al., 2011). The perception and faith of SNS users regarding the authenticity and reliability of the information source plays a crucial role in the regulation of public speech and spreading of information about moral/ethical violation topics.

Theoretical Implications

The current study offers some implications that facilitate future research on participating in discussion on moral/ethical violation topics. Public concern and discussion about moral/ethical violation topics is important for regulating negative behaviors and maintaining social justice. User discussions of this kind of behavior on Chinese SNS in recent years plays a dominant role in promoting the advancement of social institutions and governance by drawing the public attention and letting the government know some information. Therefore, investigation into this kind of behavior can lead to a deeper understanding of theories on reputation, altruistic punishment, and regulation of public speech related to social cooperation.

The current study further clarifies the relationship between identifiability and speech behavior. We confirmed the positive impact of identification and perceived anonymity on participation in discussions on moral/ethics-related events on SNS. This lays a foundation for further exploration of the influence of online user identification on behavior. In the current study, unlike previous studies, online participation in moral/ethical violation topics in cyberspace is divided into commenting and information sharing, which are affected differently by the perception of risks in discussions on moral/ethical violation topic. This means that comments on specific events might not necessarily be seen by a user's social contacts, but information sharing enables user opinions to be seen by a wide range of social contacts. Our research offers a new perspective for viewing the differences between them.

Practical Implications

The results have practical implications for policy makers, content moderators, and operators of online platforms. Online identification and anonymity perception influence participation in and speech about moral/ethical violation topics, which can have a significant reference value for identity management and internet governance. A network identity policy needs to take into account the reliability of user authentication mechanisms as well as user perception of privacy. This kind of network environment encourages user participation in discussions on sensitive topics. Policy makers should also note that the sense of risk does not necessarily inhibit behavior. In the current study, perceived risk is found to encourage user participation in discussions about moral/ethical violation topics. Successful information efforts require commitment, trust, cooperation, and a clear sense of value (Goodwin et al., 2015). Correct information values are conducive to promoting the sharing and exchange of information.

Besides, the reliability of the topics is very important to the users' participation in commenting and information sharing. It is necessary to guarantee the reliability of content and information in cyberspace for the network operators and network information providers. It is necessary to regulate the information sources and provide information authentication mechanisms to identify and eliminate false information and rumors and standardize the expression form of information, etc., to promote the discussion and spread of topics.

Limitations and Future Directions

The current study has the following limitations, which open up some avenues for further research.

First, the research study and survey participants only used Weibo and WeChat in China, neglecting the difference embedded in contexts across China and the West. A trial study aimed to what was being studied, which needs further confirmation. Future research is needed to examine how and to what extent contextual and cultural differences affect the research questions and model. Second, the research provides no empirical support for H3b, which means we failed to confirm a positive impact on users from perceived risk about their sharing intention in discussing moral/ethical issues on Chinese SNS. Although differences between posting comments and sharing information are indicated in our discussion, the reasons and mechanisms should be further explored. Third, although the sample size meets the requirements of PLS SEM research with a degree of representativeness to some extent, the limitations of the study caused by the small and non-representative sample still remain; the size and representativeness of the sample need to be expanded in future research, which will contribute to the generalizability of the findings. In addition, the research sample failed to properly consider differences in age, gender, vocation, economic status, education level, and SNS use; therefore, more variables should be taken into consideration in a future study to enhance the representativeness of the research sample.

People's participation in discussion on violations of morality and ethics on the internet is taken as a crucial form of public collective activity in the criticism and supervision of society. One of the doubts about the internet real-name system is whether identifiability will impede such power. In the current study, empirical evidence was provided on the positive effects of identification, perceived anonymity, risk perception, and information credibility on users' participation in discussions on unethical topics in Chinese SNS, with a view to providing a reference point for subsequent academic studies, information management of SNS, and the governance of society.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author/s.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent from the patients/participants or patients/participants legal guardian/next of kin was not required to participate in this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

XC performed the theory analysis and design, and contributed to drafting the manuscript. CH analyzed the data and improved the empirical analysis. YC collected data and improved the conclusions. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the China National Social Science Fund Project, named Research on Network Anonymity and Network Society Governance (Grant Number: 15CSH017), Program for Excellent Young Talents, Yunnan University, and Youth talent project of Yunnan Province, named Research on network comprehensive information management (No. C176220200).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2020.535605/full#supplementary-material

REFERENCES

Adams, J. (2003). "Risk and morality: three framing devices," in *Risk and morality*, eds Edn, eds V. E. Richard and D. Aaron (London: University of Toronto Press), 87–106. doi: 10.3138/9781442679382-006

Akdeniz, Y. (2002). Anonymity, democracy, and cyberspace. *Soc. Res.* 69, 223–237.

Bandura, A. (1977). Self-efficacy: toward a unifying theory of behavioral change. *Psychol. Rev.* 84:191. doi: 10.1037/0033-295x.84.2.191

Bodle, R. (2013). The ethics of online anonymity or Zuckerberg vs. *Moot. ACM SIGCAS Comput. Soc.* 43, 22–35. doi: 10.1145/2505414.2505417

Boeyink, D. E., and Borden, S. L. (2010). *Making hard choices in journalism ethics: Cases and practice*. Abingdon: Routledge.

Borden, S. L. (2007). *Journalism as practice: MacIntyre, virtue ethics and the press*. United Kingdom: Ashgate Publishing, Ltd.

Borden, S. L. (2016). Aristotelian casuistry: Getting into the thick of global media ethics. *Commun. Theor.* 26, 329–347. doi: 10.1111/comt.12085

Boulianne, S. (2015). Social media use and participation: A meta-analysis of current research. *Inform. Commun. Soc.* 18, 524–538. doi: 10.1080/1369118x.2015.1008542

Boyd, R., Gintis, H., and Bowles, S. (2010). Coordinated punishment of defectors sustains cooperation and can proliferate when rare. *Science* 328, 617–620. doi: 10.1126/science.1183665

Brazier, F., Oskamp, A., Prins, C., Schellekens, M., and Wijngaards, N. (2004). Anonymity and software agents: an interdisciplinary challenge. *Artif. Intell. Law* 12, 137–157. doi: 10.1007/s10506-004-6488-5

Brennan, M., Afroz, S., and Greenstadt, R. (2012). Adversarial stylometry: Circumventing authorship recognition to preserve privacy and anonymity. *ACM Trans. Inf. Syst. Secur.* 15:12.

Chang, E. C., and Tseng, Y. F. (2013). Research note: E-store image, perceived value and perceived risk. *J. Bus. Res.* 66, 864–870. doi: 10.1016/j.jbusres.2011.06.012

Chen, L. (2014). Chinese traditional culture and core values. *Guang Ming Daily* 16:0811.

Chen, X. (2018). Internet anonymous space: emerging order and governance logic. *Chong. Soc. Sci.* 08, 26–34.

Chen, X., Fang, S., Li, Y., and Wang, H. (2019a). Does Identification Influence Continuous E-Commerce Consumption? The Mediating Role of Intrinsic Motivations. *Sustainability* 11:1944. doi: 10.3390/su11071944

Chen, X., Sun, M., Wu, D., and Song, X. Y. (2019b). Information-Sharing Behavior on WeChat Moments: The Role of Anonymity, Familiarity, and Intrinsic Motivation. *Front. Psychol.* 10:2540.

Chen, X., and Li, G. (2013). Essence, evolution and freedom of expression of cyber real name system. *Xinhua Digest* 20, 132–135.

Chen, X., Li, G., Hu, Y., and Li, Y. (2016). How Anonymity Influence Self-Disclosure Tendency on Sina Weibo: An Empirical Study. *Anthropologist* 26, 217–226. doi: 10.1080/09720073.2016.11892151

Cherry, J., and Fraedrich, J. (2002). Perceived risk, moral philosophy and marketing ethics: mediating influences on sales managers' ethical decision-making. *J. Bus. Res.* 55, 951–962. doi: 10.1016/s0148-2963(00)00215-0

Cheung, C., Lee, Z. W., and Chan, T. K. (2015). Self-disclosure in social networking sites the role of perceived cost, perceived benefits and social influence. *Internet Res.* 25:279. doi: 10.1108/intr-09-2013-0192

Cheung, C. M., and Lee, M. K. (2012). What drives consumers to spread electronic word of mouth in online consumer-opinion platforms. *Decis. Supp. Syst.* 53, 218–225. doi: 10.1016/j.dss.2012.01.015

Christopherson, K. M. (2007). The positive and negative implications of anonymity in Internet social interactions:"On the Internet, Nobody Knows You're a Dog". *Comput. Hum. Behav.* 23, 3038–3056. doi: 10.1016/j.chb.2006.09.001

Chung, N., Nam, K., and Koo, C. (2016). Examining information sharing in social networking communities: applying theories of social capital and attachment. *Telemat. Inform.* 33, 77–91. doi: 10.1016/j.tele.2015.05.005

Cyranoski, D., and Ledford, H. (2018). Genome-edited baby claim provokes international outcry. *Nature* 563, 607–608. doi: 10.1038/d41586-018-07545-0

Davenport, D. (2002). Anonymity on the Internet: why the price may be too high. *Commun. ACM* 45, 33–35. doi: 10.1145/505248.505267

De Quervain, D. J., Fischbacher, U., Treyer, V., and Schellhammer, M. (2004). The neural basis of altruistic punishment. *Science* 305:1254. doi: 10.1126/science.1100735

Douglas, M. (2013). *Risk and acceptability*. United Kingdom: Routledge.

Earle, T. C., and Siegrist, M. (2006). Morality Information, Performance Information, and the Distinction Between Trust and Confidence. *J. Appl. Soc. Psychol.* 36, 383–416. doi: 10.1111/j.0021-9029.2006.00012.x

Elliott, D. (ed.) (1986). *Foundations for news media responsibility. Responsible journalism*. Thousand Oaks, CA: Sage, 32–44.

Ericson, R. V., and Doyle, A. eds (2003). *Risk and morality*. Canada: University of Toronto Press.

Featherman, M. S., and Pavlou, P. A. (2003). Predicting e-services adoption: a perceived risk facets perspective. *Int. J. Hum. Comput. Stud.* 59, 451–474. doi: 10.1016/s1071-5819(03)00111-3

Fehr, E., and Fischbacher, U. (2004). Third-party punishment and social norms. *Evol. Hum. Behav.* 25, 63–87. doi: 10.1016/s1090-5138(04)00005-4

Fehr, E., and Gächter, S. (2002). Altruistic punishment in humans. *Nature* 415, 137–140. doi: 10.1038/415137a

Fornell, C., and Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *J. Market. Res.* 18, 39–50. doi: 10.1177/002224378101800104

Fox, J., Cruz, C., and Lee, J. Y. (2015). Perpetuating online sexism offline: Anonymity, interactivity, and the effects of sexist hashtags on social media. *Comput. Hum. Behav.* 52, 436–442. doi: 10.1016/j.chb.2015.06.024

Fox, J., and Moreland, J. J. (2015). The dark side of social networking sites: an exploration of the relational and psychological stressors associated with Facebook use and affordances. *Comp. Hum. Behav.* 45, 168–176. doi: 10.1016/j.chb.2014.11.083

Froomkin, A. M. (2015). From Anonymity to Identification. *J. Self Regul. Regul.* 1:120.

Gerlach, J., Widjaja, T., and Buxmann, P. (2015). Handle with care: How online social network providers' privacy policies impact users' information sharing behavior. *J. Strat. Inform. Sys.* 24, 33–43. doi: 10.1016/j.jsis.2014.09.001

Gintis, H., Bowles, S., Boyd, R. T., and Fehr, E. (2005). *Moral sentiments and material interests: The foundations of cooperation in economic life*. Cambridge: MIT press.

Gist, M. E., and Mitchell, T. R. (1992). Self-efficacy: A theoretical analysis of its determinants and malleability. *Acad. Manag. Rev.* 17, 183–211. doi: 10.2307/258770

Global Web Index. (2018). *Social Global Web Index's Flagship Report on the Latest Trends in Social Media*. London: Global Web Index.

Goldsmith, R. E., and Horowitz, D. (2006). Measuring motivations for online opinion seeking. *J. Interact. Advertis.* 6, 2–14. doi: 10.1080/15252019.2006.10722114

Goodwin, C., Nicholas, J. P., Bryant, J., Ciglic, K., Kleiner, A., Kutterer, C., et al. (2015). *A framework for cybersecurity information sharing and risk reduction*. New Mexico: Microsoft Corporation.

Guan, M., and So, J. (2016). Influence of social identity on self-efficacy beliefs through perceived social support: A social identity theory perspective. *Commun. Stud.* 67, 588–604. doi: 10.1080/10510974.2016.1239645

Ha, S., and Ahn, J. (2011). ""Why Are You Sharing Others' Tweets?: The Impact of Argument Quality and Source Credibility on Information Sharing Behavior," in *Proceedings of the International Conference on Information Systems, ICIS 2011*, (China: ICIS), 4.

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., and Tatham, R. L. (2006). Multivariate data analysis 6th Edition. Pearson Prentice Hall. New Jersey. humans: Critique and reformulation. *J. Abnor. Psychol.* 87, 49–74.

Hajli, N., and Lin, X. (2016). Exploring the security of information sharing on social networking sites: The role of perceived control of information. *J. Bus. Ethics* 133, 111–123. doi: 10.1007/s10551-014-2346-x

Henseler, J., Dijkstra, T. K., Sarstedt, M., Ringle, C. M., Diamantopoulos, A., Straub, D. W., et al. (2014). Common beliefs and reality about PLS: Comments on Rönkkö and Evermann (2013). *Organizat. Res. Methods* 17, 182–209. doi: 10.1177/1094428114526928

Hertzum, M., Andersen, H. H., Andersen, V., and Hansen, C. B. (2002). Trust in information sources: seeking information from people, documents, and virtual agents. *Interact. Comput.* 14, 575–599. doi: 10.1016/s0953-5438(02)00023-1

Hite, D. M., Voelker, T., and Robertson, A. (2014). Measuring perceived anonymity: The development of a context independent instrument. *J. Methods Measur. Soc. Sci.* 5, 22–39. doi: 10.2458/v5i1.18305

Hong, I. B. (2015). Understanding the consumer's online merchant selection process: The roles of product involvement, perceived risk, and trust expectation. *Int. J. Inform. Manag.* 35, 322–336. doi: 10.1016/j.ijinfomgt.2015.01.003

Horst, M., Kuttschreuter, M., and Gutteling, J. M. (2007). Perceived usefulness, personal experiences, risk perception and trust as determinants of adoption of e-government services in The Netherlands. *Comput. Hum. Behav.* 23, 1838–1852. doi: 10.1016/j.chb.2005.11.003

Hou, J., Ndasauka, Y., Pan, X., Chen, S., Xu, F., and Zhang, X. (2018). Weibo or WeChat? Assessing preference for social networking sites and role of personality traits and psychological factors. *Front. Psychol.* 9:545.

Hsieh, A. Y., and Luarn, P. (2014). Speech or silence: the effect of user anonymity and member familiarity on the willingness to express opinions in virtual communities. *Online Inform. Rev.* 38, 881–895. doi: 10.1108/oir-03-2014-0076

Huang, X. C., and Zhang, Q. L. (2010). Discussion on the real name system in the virtual society from the real name system in the real society. *J. Southern Stud.* 8, 60–62.

Jang, Y. J., Kim, H. W., and Jung, Y. (2016). A mixed methods approach to the posting of benevolent comments online. *Int. J. Inform. Manag.* 36, 414–424. doi: 10.1016/j.ijinfomgt.2016.02.001

Jardine, E. (2015). "The Dark Web dilemma: Tor, anonymity and online policing," in *Global Commission on Internet Governance Paper Series, (21)*, (Waterloo: Centre for International Governance Innovation).

Jenkins, R. (2014). *Social Identity*. London, UK: Routledge, 2014.

Jessup, L. M., Connolly, T., and Galegher, J. (1990). The effects of anonymity on GDSS group process with an idea-generating task. *MIS Q.* 14, 313–321. doi: 10.2307/248893

Joinson, A. N. (2001). Self-disclosure in computer-mediated communication: The role of self-awareness and visual anonymity. *Eur. J. Soc. Psychol.* 31, 177–192. doi: 10.1002/ejsp.36

Joinson, A. N. (2007). *Oxford Handbook of Internet Psychology*. Oxford, UK: Oxford University Press, 2007.

Kang, R., Brown, S., and Kiesler, S. (2013). "Why do people seek anonymity on the internet? Informing policy and design," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, New York, NY, 2657–2666.

Kelton, K., Fleischmann, K. R., and Wallace, W. A. (2008). Trust in digital information. *J. Am. Soc. Inform. Sci. Technol.* 59, 363–374.

Kim, W., Jeong, O. R., Kim, C., and So, J. (2011). The dark side of the Internet: Attacks, costs and responses. *Inform. Sys.* 36, 675–705. doi: 10.1016/j.is.2010.11.003

Kiousis, S. (2001).). Public trust or mistrust? Perceptions of media credibility in the information age. *Mass Commun. Soc.* 4, 381–403. doi: 10.1207/s15327825mcs0404_4

Koch, N. S., and Emrey, J. A. (2001). The Internet and opinion measurement: Surveying marginalized populations. *Soc. Sci. Q.* 82, 131–138. doi: 10.1111/0038-4941.00012

Krausová, A. (2009). Identification in Cyberspace. *Masaryk Unive. J. Law Technol.* 2, 83–95.

Kuersten, A., and Wexler, A. (2019). Ten ways in which He Jiankui violated ethics. *Nat. Biotechnol.* 37:19. doi: 10.1038/nbt.4337

Kwon, S. (2020). Understanding user participation from the perspective of psychological ownership: The moderating role of social distance. *Comput. Hum. Behav.* 105:106207. doi: 10.1016/j.chb.2019.106207

Lapidot-Lefler, N., and Barak, A. (2012). Effects of anonymity, invisibility, and lack of eye-contact on toxic online disinhibition. *Comput. Hum. Behav.* 28, 434–443. doi: 10.1016/j.chb.2011.10.014

Levontin, L., and Yom-Tov, E. (2017). Negative Self-Disclosure on the Web: The Role of Guilt Relief. *Front. Psychol.* 8:1068.

Li, R., and Suh, A. (2015). Factors influencing information credibility on social media platforms: evidence from Facebook pages. *Proc. Comput. Sci.* 72, 314–328. doi: 10.1016/j.procs.2015.12.146

Liebman, B. L. (2011). The media and the courts: towards competitive supervision? *China Q.* 208, 833–850. doi: 10.1017/s0305741011001020

Lin, R., and Utz, S. (2017). Self-disclosure on sns: do disclosure intimacy and narrativity influence interpersonal closeness and social attraction? *Comput. Hum. Behav.* 70, 426–436. doi: 10.1016/j.chb.2017.01.012

Lin, Y. Q. (2010). *Debate on advantages and disadvantages of real name system on the Internet: advantages far outweigh disadvantages*. Beijing: People's daily.

Lipschultz, J. (2018). *Free expression in the age of the Internet: Social and legal boundaries*. London: Routledge.

Liu, J. M. (2013). Coercion and tolerance of real name system. *Contemp. Commun.* 01, 41–42.

Lucassen, T., and Schraagen, J. M. (2010). "Trust in wikipedia: how users trust information from an unknown source," in *Proceedings of the 4th workshop on Information credibility*, (New York: ACM), 19–26.

Lupton, D. (1993). Risk as moral danger: the social and political functions of risk discourse in public health. *Int. J. Health Serv.* 23, 425–435. doi: 10.2190/16ay-e2gc-dfld-51x2

Lyons, B., and Henderson, K. (2005). Opinion leadership in a computer-mediated environment. *J. Consum. Behav. Int. Res. Rev.* 4, 319–329. doi: 10.1002/cb.22

Marsh, H. W., Hau, K. T., and Grayson, D. (2005). *Goodness of fit in structural equation models*. USA: Lawrence Erlbaum Associates Publishers, 2005.

Martins, C., Oliveira, T., and Popović, A. (2014). Understanding the Internet banking adoption: a unified theory of acceptance and use of technology and perceived risk application. *Int. J. Inform. Manage.* 34, 1–13. doi: 10.1016/j.ijinfomgt.2013.06.002

Marx, G. T. (1999). What's in a Name? Some Reflections on the Sociology of Anonymity. *Inform. Soc.* 15, 99–112. doi: 10.1080/019722499128565

McKnight, D. H. (2005). Trust in information technology. *Blackwell Encycl. Manag.* 7, 329–333.

Mitchell, V. W. (1999). Consumer perceived risk: conceptualisations and models. *Eur. J. Market.* 33, 163–195. doi: 10.1108/03090569910249229

Nicolaou, A. I., and McKnight, D. H. (2006). Perceived information quality in data exchanges: effects on risk, trust, and intention to use. *Inform. Sys. Res.* 17, 332–351. doi: 10.1287/isre.1060.0103

Nissenbaum, H. (1999). The meaning of anonymity in an information age. *Inform. Soc.* 15, 141–144. doi: 10.1080/019722499128592

Normile, D. (2018). Shock greets claim of CRISPR-edited babies. *Science* 362, 978–979. doi: 10.1126/science.362.6418.978

Özer, Ö, Zheng, Y., and Chen, K. Y. (2011). Trust in forecast information sharing. *Manag. Sci.* 57, 1111–1137. doi: 10.1287/mnsc.1110.1334

Pan, L. Y., and Chiou, J. S. (2011). How much can you trust online information? Cues for perceived trustworthiness of consumer-generated online information. *J. Interact. Market.* 25, 67–74. doi: 10.1016/j.intmar.2011.01.002

Peter, J. P., and Ryan, M. J. (1976). An investigation of perceived risk at the brand level. *J. Market. Res.* 13, 184–188. doi: 10.1177/002224377601300210

Rainie, L., Kiesler, S., Kang, R., Madden, M., Duggan, M., Brown, S., et al. (2013). Anonymity, privacy, and security online. *Pew Res. Center* 5, 1–35.

Rauchfleisch, A., and Schäfer, M. S. (2015). Multiple public spheres of Weibo: A typology of forms and potentials of online public spheres in China. *Inform. Commun. Soc.* 18, 139–155. doi: 10.1080/1369118x.2014.940364

Reinig, B. A., and Mejias, R. J. (2004). The efects of national culture and anonymity on flaming and criticalness in GSS-supported discussions. *Small Group Res.* 35, 698–723. doi: 10.1177/1046496404266773

Repnikova, M. (2017). Media openings and political transitions: Glasnost versus Yulun Jiandu. *Probl. Post Commun.* 64, 141–151. doi: 10.1080/10758216.2017.1307118

Ringle, C. M., Wende, S., and Becker, J. M. (2014). SmartPLS 3. Hamburg: SmartPLS. *Acad. Manag. Rev.* 9, 419–445.

Salanova, M., Llorens, S., and Cifre, E. (2013). The dark side of technologies: Technostress among users of information and communication technologies. *Int. J. Psychol.* 48, 422–436. doi: 10.1080/00207594.2012.680460

Sander, T., and Phoey, L. T. (2014). *SmartPLS for the human resources field to evaluate a model. Conference: New Challenges of Economic and Business DevelopmentAt*. Latvia: University of Latvia, 346–358.

Scott, S. V., and Orlikowski, W. J. (2014). Entanglements in Practice: Performing Anonymity Through Social Media. *MIS Q.* 38, 873–893. doi: 10.25300/misq/2014/38.3.11

Shah, D. V., Cho, J., Eveland, W. P. Jr., and Kwak, N. (2005). Information and expression in a digital age: Modeling Internet effects on civic participation. *Commun. Res.* 32, 531–565. doi: 10.1177/0093650205279209

Shen, F., Wang, N., Guo, Z., and Guo, L. (2009). Online network size, efficacy, and opinion expression: Assessing the impacts of Internet use in China. *Int. J. Publ. Opin. Res.* 21, 451–476. doi: 10.1093/ijpor/edp046

Slovic, P. (1987). Perception of risk. *Science* 236, 280–285.

Slovic, P. (1993). Perceived risk, trust, and democracy. *Risk Anal.* 13, 675–682. doi: 10.1111/j.1539-6924.1993.tb01329.x

Snoj, B., Pisnik Korda, A., and Mumel, D. (2004). The relationships among perceived quality, perceived risk and perceived product value. *J. Prod. Brand Manag.* 13, 156–167. doi: 10.1108/10610420410538050

Spears, R. (2017). Social identity model of deindividuation effects. *Int. Encycl. Med. Effec.* 4, 1–9. doi: 10.1002/9781118783764.wbieme0091

Stroud, S. R. (2014). The dark side of the online self: A pragmatist critique of the growing plague of revenge porn. *J. Mass Med. Ethics* 29, 168–183. doi: 10.1080/08900523.2014.917976

Sullivan, J. (2014). China's Weibo: Is faster different? *New Med. Soc.* 16, 24–37. doi: 10.1177/1461444812472966

Valenzuela, S. (2013). Unpacking the use of social media for protest behavior: The roles of information, opinion expression, and activism. *Am. Behav. Sci.* 57, 920–942. doi: 10.1177/0002764213479375

Van den Eijnden, R. J., Meerkerk, G. J., Vermulst, A. A., Spijkerman, R., and Engels, R. C. (2008). Online communication, compulsive Internet use, and psychosocial well-being among adolescents: A longitudinal study. *Dev. Psychol.* 44:655. doi: 10.1037/0012-1649.44.3.655

van Winsen, F., de Mey, Y., Lauwers, L., Van Passel, S., Vancauteren, M., and Wauters, E. (2016). Determinants of risk behaviour: effects of perceived risks and risk attitude on farmer's adoption of risk management strategies. *J. Risk Res.* 19, 56–78. doi: 10.1080/13669877.2014.940597

Vilanova, F., Beria, F. M., Costa, ÂB., and Koller, S. H. (2017). Deindividuation: From Le Bon to the social identity model of deindividuation effects. *Cogent Psychol.* 4:1308104.

Weeks, B. E., Ardèvol-Abreu, A., and Gil de Zúñiga, H. (2017). Online influence? Social media use, opinion leadership, and political persuasion. *Int. J. Publ. Opin. Res.* 29, 214–239.

Wen, Z. L., and Ye, B. J. (2014). Analyses of mediating effects: the development of methods and models. *Adv. Psychol. Sci.* 22, 731–745. doi: 10.3724/sp.j.1042.2014.00731

Wright, M. F. (2014). Predictors of anonymous cyber aggression: the role of adolescents' beliefs about anonymity, aggression, and the permanency of digital content. *Cyberpsychol. Behav. Soc. Network.* 17, 431–438. doi: 10.1089/cyber.2013.0457

Yan, Z., Wang, T., Chen, Y., and Zhang, H. (2016). Knowledge sharing in online health communities: a social exchange theory perspective. *Inf. Manag.* 53, 643–653. doi: 10.1016/j.im.2016.02.001

Yoon, C., and Rolland, E. (2012). Knowledge-sharing in virtual communities: familiarity, anonymity and self-determination theory. *Behav. Inform. Technol.* 31, 1133–1143. doi: 10.1080/0144929x.2012.702355

Yu, G. M. (2017). Key elements and operation on network public opinion governance. *News Writ.* 1, 10–13.

Yue, H. Y., Pang, W. M., Tam, C. Y., and Fan, C. W. N. (2017). "Does Spending More Time on Facebook Makes Users Engage in Politics?," in *2017 IEEE International Symposium on Multimedia (ISM)*, (Berling: IEEE), 426–431.

Zhai, X. W. (1999). Personal status: a concept and its analytical framework: the real construction of China's daily society. *Chin. Soc. Sci.* 4, 144–157.

Zhang, F., and Lu, Y. (2010). Thinking about power consciousness based on the dispute of real name system on the Internet. *Seek. Truth* 11, 72–75.

Zhang, L., Zhao, J., and Xu, K. (2015). Who creates trends in online social media: The crowd or opinion leaders? *J. Comput. Med. Commun.* 21, 1–16. doi: 10.1111/jcc4.12145

Zimmer, J. C., Arsal, R. E., Al-Marzouq, M., and Grover, V. (2010). Investigating online information disclosure: Effects of information relevance, trust and rbisk. *Inform. Manag.* 58, 56–70.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Psychology

ORIGINAL RESEARCH
published: 04 November 2020
doi: 10.3389/fpsyg.2020.546546

Taking Risks With Cybersecurity: Using Knowledge and Personal Characteristics to Predict Self-Reported Cybersecurity Behaviors

Shelia M. Kennison[1]* *and Eric Chan-Tin*[1,2]

[1] Department of Psychology, Oklahoma State University, Stillwater, OK, United States, [2] Department of Computer Science, Loyola University Chicago, Chicago, IL, United States

Individuals' use of insecure cybersecurity behaviors, including the use of weak passwords, is a leading contributor to cybersecurity breaches. While training individuals on best practices in cybersecurity continues to be implemented, prior research has found that training people in the use of secure passwords has not proven to be effective. Developing profiles of individual who are likely to become victims of password hacking, phishing scams, and other types of breaches would be useful, as they could be used to identify individuals with the highest likelihood of engaging in insecure cybersecurity behaviors. The present research tested the hypothesis that in addition to self-reported cybersecurity knowledge, personal characteristics, such as personality traits and general risk-taking behavior not related to technology use, can predict individual differences in cybersecurity behaviors, as measured by self-report. Our hypothesis was confirmed in a large study involving 325 undergraduates. Participants provided information about their self-reported risky cybersecurity behaviors (e.g., using non-secure Wi-Fi, not logging out of accounts on shared computers, etc.), self-reported knowledge about strong/weak passwords, Big Five personality traits (i.e., extraversion, conscientiousness, agreeableness, openness, and mood instability), sensation-seeking personality traits, and general risk-taking unrelated to using technology. The results of a hierarchical regression indicated that 34% of risky cybersecurity behavior was significantly predicted by the combination of self-reported knowledge about strong/weak passwords, personality traits, and risk-taking in daily life. The results suggest that victim profiles should take into account individual differences in personality and general risk-taking in domains unrelated to cybersecurity in addition to cybersecurity knowledge.

Keywords: risk-taking, cybersecurity, passwords, personality, DOSPERT

OPEN ACCESS

Edited by:
Richard Keith Wortley,
University College London,
United Kingdom

Reviewed by:
Stuart McKelvie,
Bishop's University, Canada
Burak Merdenyan,
University of York, United Kingdom

****Correspondence:***
Shelia M. Kennison
shelia.kennison@okstate.edu

Specialty section:
This article was submitted to
Cognition,
a section of the journal
Frontiers in Psychology

Received: *28 March 2020*
Accepted: *13 October 2020*
Published: *04 November 2020*

Citation:
Kennison SM and Chan-Tin E
(2020) Taking Risks With
Cybersecurity: Using Knowledge
and Personal Characteristics
to Predict Self-Reported
Cybersecurity Behaviors.
Front. Psychol. 11:546546.
doi: 10.3389/fpsyg.2020.546546

INTRODUCTION

The average American has little awareness of cybersecurity issues, despite the fact that the majority have been affected by some type of security breach (Pew Research Center, 2017). Research has documented that using weak passwords and re-using passwords for multiple accounts is common (Gaw and Felten, 2006; Florencio and Herley, 2007; Grawemeyer and Johnson, 2011). Recent research has explored strategies for reducing computer users' vulnerabilities by educating them about the dangers of risky cybersecurity behaviors, such as choosing weak passwords (Farcasin and Chan-Tin, 2015) and re-using passwords (Stobert and Biddle, 2014). Research has shown that educating people on security best practices through trainings may not be effective (Riley, 2006; Lorenz et al., 2013). Studies have shown that those with knowledge about password security will, nonetheless, use weak passwords and/or re-use passwords in their daily lives (Riley, 2006; Notoatmodjo and Thomborson, 2009). Nevertheless, continued efforts to develop and to test the effectiveness of training curriculum are warranted (Bryant and Campbell, 2006; Taylor-Jackson et al., 2020). Few studies have investigated whether personality traits predict knowledge and cybersecurity behaviors (e.g., Whitty et al., 2015). The focus of the present research was to determine whether risky cybersecurity behavior could be predicted from a combination of password security knowledge and personal characteristics, such as personality traits and general risk-taking in daily life.

Case studies of cybersecurity breaches have shown that humans, rather than technology, are the *weakest link*, responsible for risky cybersecurity behaviors that provide access points for cybersecurity attacks (Mitnick, 2003; Pew Research Center, 2017; cf. Adams and Sasse, 1999). Numerous security breaches have involved the use of weak passwords. Some examples include the credit report company Equifax (Wang and Johnson, 2018), the retailer Target (Plachkinova and Maurer, 2019), and an American university (Ayyagari and Tyks, 2012). An increasing number of platforms are implementing requirements for users to use stronger passwords (i.e., with a combination of numbers, lowercase and uppercase letters, and other symbols or a passphrase); however, security vulnerability remains when users use the same password for multiple accounts (Thomas et al., 2017).

Numerous studies have explored the effectiveness of cybersecurity training to increase users' knowledge about best cybersecurity practices and to decrease the use of risky cybersecurity behaviors in daily life (Ferguson, 2005; McCrohan et al., 2010; Peker et al., 2016; see for review, Proctor, 2016). A prevalent view is that institutions should not rely solely on cybersecurity training, because in the past, it has not been shown to be effective (Ferguson, 2005; Bada et al., 2019). There is also the recognition that regardless of how much training an institution carries out, a security breach can occur from a small number of individuals' risky cybersecurity behaviors. In a large sample of student participants, Riley (2006) showed that individuals may use weak passwords despite reporting that they knew that such passwords were not the most secure. Nevertheless, several recent studies have demonstrated some positive benefits of training (McCrohan et al., 2010; Peker et al., 2016). The trainings that have shown benefits have focused on providing individuals with knowledge about cybersecurity threats and best cybersecurity practices. Adams and Sasse (1999) suggested that users' lack of knowledge about cybersecurity and their perceptions of insecure practices as low-risk may be due to inadequate communication to users from the relevant institutional entities.

The present research examined the possibility that it may be possible to predict individual differences in risky cybersecurity behaviors using personal characteristics, such as knowledge about password security, personality traits, and personality-related behaviors. Prior research has found that men report higher levels of knowledge about cybersecurity than women (Cain et al., 2018) and also higher levels of risky cybersecurity behavior (Anwar et al., 2017). Numerous studies have examined the relationships among Big Five personality traits (i.e., extraversion, agreeableness, openness, conscientiousness, and mood instability) and cybersecurity behavior (McBride et al., 2012; Tamrakar et al., 2016; McCormac et al., 2017; Russell et al., 2017; Alohali et al., 2018; Shappie et al., 2019). These studies have looked at the relationship between personality traits and cybersecurity behavior. Big five personality traits have been described as universal (Yamagata et al., 2006; cf. Gurven et al., 2013) and stable across the lifespan (Conley, 1985). Tamrakar et al. (2016) created a simulation tool to measure the relationship between personality traits and cyber behaviors. Russell et al. (2017) studied how people engaged in secure cybersecurity behaviors are more positive. They also found that secure cybersecurity behaviors are linked to emptiness and meaningless while greater use of insecure cybersecurity behaviors are related to lower conscientiousness and higher levels of aggressive behavior.

Some studies are slightly contradicting. Shappie et al. (2019) showed that conscientiousness, agreeableness, and openness were significantly associated with cybersecurity behaviors. In contrast, Alohali et al. (2018) showed that conscientiousness is negatively correlated with cybersecurity behaviors. The Human Aspects of Information Security Questionnaire (HAIS-Q) was utilized in McCormac et al. (2017); they have shown that conscientiousness, agreeableness, emotional stability, and risk-taking propensity significantly explained the variance in individuals' score, while age and gender did not. While most papers recruited participants from schools (McBride et al., 2012), recruited IT practitioners and looked at how likely these practitioners are to violate cybersecurity protocols based on their Big Five personality traits. In a recent study by Maraj et al. (2019), there was no relationship found between password strength and personality traits.

In addition to examining the relationships among the Big Five traits and risky cybersecurity behaviors, we also examined sensation-seeking personality traits and the extent to which individuals take risks in daily life that were unrelated to the use of technology. Sensation-seeking personality traits were first identified by Zuckerman et al. (1964, 1978; Horvath and Zuckerman, 1993; Zuckerman, 1994) and defined as the propensity to seek out new experiences with a preference toward intense experiences. Numerous studies have shown that individuals higher in sensation-seeking traits take more risks in daily life, including participating in sports (Zuckerman, 1983a), smoking, drinking and using other drugs (Zuckerman, 1983b, 1987; Zuckerman et al., 1990; Popham et al., 2011; Kennison and Messer, 2017, 2019), engaging in risky sexual behaviors (Zuckerman et al., 1976), and risky behaviors occurring during gambling (Anderson and Brown, 1984). Numerous studies suggest that sensation-seeking traits stem from individual differences in biology (see for review Roberti, 2004).

We reasoned that individuals with higher levels of sensation-seeking personality traits would engage in higher levels of risky cybersecurity behaviors and those who engage in higher levels of general risk-taking in daily life would engage in higher levels of risky cybersecurity behavior. To measure general risk-taking in daily life, we used the Domain-Specific Risk Taking (DOSPERT) scale (Blais and Weber, 2001, 2006; Weber et al., 2002; Figner and Weber, 2011), which assesses risk-taking in five domains: (a) health/safety, (b) recreational, (c) social, (d) financial, and (e) ethical. Multiple studies in which the scale was used have shown that there are significant correlations for risk-taking for the five domains, suggesting that high risk-taking in one domain predicts high risk-taking in the other domains (Kennison et al., 2016; Shou and Olney, 2020). Shou and Olney (2020) carried out a large meta-analysis using 104 samples with more than 30,000 observations and found that the five domains were intercorrelated. The health/safety domain was strongly correlated with recreational and ethical domains. The social domain was more weakly correlated with the other domains. Prior research has also observed differences in the perception of risk for men and women, with women perceiving more risk generally than men and being more risk-adverse (Gustafsod, 1998; Weber et al., 2002). Men also report engaging in risk-taking in daily life more than women (Kennison et al., 2016; see Panno et al., 2018 for review). Men also report higher levels of sensation-seeking traits than women (Kennison et al., 2016).

In this paper, we report a study that was carried out online in which we investigated how well self-reported opinion about knowledge about secure passwords, personality traits, and general risk-taking in daily life predict self-reported risky cybersecurity behaviors. Increasingly, researchers are carrying out research via the Internet (Buchanan and Smith, 1999; Gosling et al., 2004; Weigold et al., 2013; Dodou and de Winter, 2014), which leads to lower costs as staff are not needed for data entry after study completion. Internet research has been positively impacted by the increasing availability of Internet access and inexpensive survey building tools. Confidence in online research has grown due to studies that have compared data collected via the Internet and in face-to-face settings in which questionnaires were completed using pencil/paper methods and have concluded that the two data collection methods yield similar results (Gosling et al., 2004; Weigold et al., 2013). Differences in response rates, amount of missing data, and factor structure of some variables have been observed (see Weigold et al., 2013 for review). Some have suggested that participants in studies carried out via the Internet may differ in their tendency to provide socially desirable responses than participants in studies carried out in face-to-face settings (see Dodou and de Winter, 2014 for review). Dodou and de Winter (2014) carried out a meta-analysis of 51 prior research studies in which social desirability responding was compared for online and face-to-face studies. In the meta-analysis, they found that social desirability responding was similar for the two methodologies. Others have suggested that in some cases, participants may be willing to be more truthful in responding in online surveys versus studies conducted in face-to-face settings (Bailey et al., 2000).

In our study, we tested the following hypotheses: (a) higher levels of self-reported password security knowledge would be related to engaging in lower levels of risky cybersecurity behaviors (see Ferguson, 2005; McCrohan et al., 2010; Peker et al., 2016), (b) higher levels of conscientiousness will be related to lower levels of risky cybersecurity behaviors (see McCormac et al., 2017; Russell et al., 2017; Alohali et al., 2018; Shappie et al., 2019), (c) higher levels of mood instability will be related to higher levels of self-reported risky cybersecurity behavior (see McCormac et al., 2017), (d) higher levels of sensation-seeking will be related to higher levels of self-reported risky cybersecurity behaviors (cf. Whitty et al., 2015), (e) higher levels of general risk-taking behaviors will be related to higher levels of risky cybersecurity behaviors, and (f) men would engage in higher levels of risk-taking (i.e., general risk-taking and risky cybersecurity behavior) than women. We did not expect to observe significant relationships between personality traits and knowledge, as prior research has not provided evidence for these relationships and also because gaining knowledge able password security would not be expected to depend on personality traits. Knowledge is gained through communications schools or workplaces, which are experienced by people regardless of their personality traits.

MATERIALS AND METHODS

Participants

There were 325 participants (207 women, 117 men, and 1 *other*) who were taking classes in psychology or speech communications a large public university in the Midwestern region of the United States. All participants received credit that could be used for course requirements or extra credit. Participants were on average 19.46 years old (*SD* = 2.34).

Procedure and Materials

This study was carried out in accordance with the recommendations of Oklahoma State University's Institutional Review Board (IRB), which approved the protocol. After obtaining IRB approval for the study, we recruited volunteers from a research participant SONA pool in a psychology department. In the SONA recruitment description for the study, participants were told that the purpose of the study was "The purpose of this research is to investigate the relationship between password security beliefs and behaviors with personality and demographic variables." All participants gave informed consent in accordance with the Declaration of Helsinki. As recommended for all surveys conducted via the Internet (Kraut et al., 2004), the first page of our survey provided participants with information about the study and an opportunity to volunteer for the study. The research was conducted with a waiver of documentation of consent, which is common with surveys conducted over the Internet. Participants completed an online survey created using a Professional license of Surveymonkey.com. On the first page of the survey, participants viewed information about the study and instructions on how to volunteer or to decline to volunteer. All participants completed a survey in the same order. The following order was used: Big Five personality traits, sensation-seeking personality, general risk-taking in daily life, cybersecurity behavior, knowledge, and demographics. On average, participants took 37 min to complete the survey.

We assessed participants' use of risky cybersecurity practices using six items created for the present research. We considered some of the most common risk cybersecurity behaviors that would be relevant to young adults in a college setting relying on direction from prior research (Peker et al., 2016; Ramlo and Nicholas, 2020). We generated six items focused on a situation that would likely be familiar to most students on our campus. Each item addressed one cybersecurity behavior. The prior literature identified more than six problem behaviors. We chose the six problem behaviors that we believed would likely be familiar to most student on our campus and carried out a focus group of undergraduates who did not participate in the study. We confirmed from the group that the behavior would likely be familiar to most of their peers. In the survey, each item was paired with a 7-point scale (1 = *not at all likely* and 7 = *extremely likely*). The scale numbers in-between were not labeled. Each item described a practice that should be avoided. The items were: (a) *using weak passwords (e.g., pass1234)*, (b) *failing to log out of a shared computer, such as in a campus computer lab*, (c) *clicking on an unfamiliar URL link that you receive in an email*, (d) *using public unsecured Wi-Fi*, (e) *using the same password for multiple devices/applications*, and (f) *telling your password to someone at your workplace*. Items were presented in random order for each participant. We computed the mean rating for the six items with higher means reflecting higher levels of secure self-reported behavior. We observed good internal consistency for the four items (Cronbach α = 0.77, see Taber, 2018 for discussion of importance of internal consistency in psychometric measures). Nunnally (1978) suggests values above 0.70 reflect good internal consistency. Below 0.70 is viewed as questionable (George and Mallery, 2003).

We assessed participants' rating of their opinion of their own knowledge of password security using four items created for this study. Each item was paired with a 7-point scale (1 = *Strongly Disagree* and 7 = *Strongly Agree*). The scale number in-between were also labeled (i.e., 2 = *Moderately Disagree*, 3 = *Somewhat Disagree*, 4 = *Neither Disagree nor Agree*, 5 = *Somewhat Agree*, and 6 = *Moderately Agree*). The questions were: (a) *My knowledge of password security is high*, (b) *Password security practices are not something that I have learned very much about*, (c) *I know a lot about password security practices*, and (d) *My level of knowledge about real world cases where sensitive data have been stolen by hackers is fairly high*. Items were presented in random order. After reverse scoring the second item in the above sequence, a mean score for the four items was calculated. We observed good internal consistency for the four items (Cronbach α = 0.74; see Taber, 2018). The items contain some overlap. We examined correlations with subsets of the items and found similar results as when all items were used. We are reporting the results for all the items for this reason.

Sensation-seeking personality traits were assessed using Zuckerman et al. (1978) 40-item SSS-V Scale. The SSS-V is composed of four factors: (a) thrill and adventure seeking (TAS, i.e., affinity for participating in activities characterized as dangerous), (b) experience seeking (ES, i.e., interest in seeking out new experiences including unusual lifestyle practices), (c) disinhibition (DIS, i.e., affinity for out-of-control experiences, such as those that occur experiences with drugs, parties, or sexual interactions), and (d) boredom susceptibility (BS, i.e., dislike of feeling bored, including being around people who are boring). For each of the 40 items, participants viewed two statements and were asked to choose the one that best described them [e.g., (a) *I like "wild" uninhibited parties*. vs. (b) *I prefer quiet parties with good conversation*]. Prior research has demonstrated the validity of the scale (Zuckerman and Link, 1968). Prior research has shown that these factors have good internal consistency; the Cronbach alphas for the four factors: TAS (α = 0.78), DIS (α = 0.76), ES (α = 0.72), and BS (α = 0.74) (Kennison et al., 2016). In the present research, we also observed good internal consistency for the four factors with Cronbach values ranging from α = 0.72 to α = 0.78 (see Nunnally, 1978; George and Mallery, 2003; Taber, 2018).

We assessed risk-taking in daily life using Blais and Weber's (2006) 30-item DOSPERT. The scale is composed of five types of risk-taking: health (i.e., risk-taking in the form of careless as well as abuse of drugs), recreational (i.e., risk-taking when doing sports and other recreational activities), social (taking risks with social interactions, such as risky behaviors with superiors), financial (i.e., risk-taking with money), and ethical (i.e., engaging in criminal behavior as well as lying and cheating). The 30-items are specific behaviors, and participants rate on a 7-point rating scale ranging from 1 (*Extremely Unlikely*) to 7 (*Extremely Likely*) how likely they are to engage in the behaviors. The scale number in-between were also labeled (i.e., 2 = *moderately unlikely*, 3 = *somewhat unlikely*, 4 = *neither unlikely nor likely*, 5 = *somewhat likely*, and 6 = *moderately likely*). Prior research has demonstrated the validity of the scale (Frey et al., 2017) and has shown that the five domains have good internal consistency: (a) health (α = 0.76), (b) recreational (α = 0.84), (c) social (α = 0.71), (d) financial (α = 0.84), and ethical (α = 0.83) (Kennison et al., 2016). In the present research, we also observed good internal consistency for the five domains with Cronbach alphas ranging from α = 0.71 to α = 0.84 (see Nunnally, 1978; George and Mallery, 2003; Taber, 2018).

We asked participants about their Big 5 personality traits (i.e., extraversion, agreeableness, conscientiousness, openness, and mood instability) using Saucier's (1994) Mini-Marker measure, which contains 40 adjectives (i.e., 8 for each trait). Participants are asked how accurate each adjective is in describing them using a 9-point scale (1 = *extremely inaccurate* and 9 = *extremely accurate*). The scale number in-between were also labeled (i.e., 2 = *very inaccurate*, 3 = *moderately inaccurate*, 4 = *slightly inaccurate*, 5 = *neither accurate nor inaccurate*, 6 = *slightly accurate*, 7 = *moderately accurate*, and 8 = *very accurate*). After reverse scoring when appropriate, we calculated the average rating for the eight adjectives for each trait. The validity of the measure has been demonstrated in prior research (Dwight et al., 1998). The measure is associated with high internal consistency (Cronbach alphas between from 0.76 to 0.86, Mooradian and Nezlek, 1996). We also observed high internal consistency in the present study (Cronbach alphas between $\alpha = 0.69$ and $\alpha = 0.82$; see Nunnally, 1978; George and Mallery, 2003; Taber, 2018).

One question was included as an attention check, assessing participants' attention to the survey with a 5-point response scale (1 = *strongly disagree*, 2 = *slightly disagree*, 3 = *neither disagree nor agree*, 4 = *slightly agree*, and 5 = *strongly agree*). Each option was listed on a separate line in multiple choice format. The question text was as follows: *Sometimes researchers include a question to determine if the participant is paying adequate attention while completing the survey. In order to show us that you are paying attention please select the third option as the response to this question*.

RESULTS

The dataset including participants' responses was screened to detect any participants who incorrectly responded to the attention check question. Thirty-three participants were removed from the dataset. The resulting dataset contained data from 292 participants (186 women, 105 men, and 1 who selected *other* for gender). **Table 1** displays means, standard deviations, and Pearson's *r* product-moment correlations for the variables that we measured in the study. Prior to conducting the correlations, we examined ranges for all variables and found no indication that there was restriction of range. The results indicated support or partial support for the four hypotheses: (a) higher levels of self-reported password security knowledge were related to lower levels of self-reported risky cybersecurity behaviors, (b) higher levels of sensation-seeking were related to higher levels of self-reported risky cybersecurity behaviors for women, but not men, (c) higher levels of mood instability were related to higher levels of self-reported risky cybersecurity behaviors, (d) higher levels of general risk-taking behaviors were related to higher levels of self-reported risky cybersecurity behaviors, (e) higher levels of conscientiousness were related to lower levels of risky cybersecurity behaviors for women, but not men, and (f) men reported engaging in higher levels of general risk-taking than women, $t(286) = -5.54$, $p < 0.001$, and η^2 = 0.41l, but there was no significant difference in self-reported risky cybersecurity behavior for men and women. Contrary to expectations, we found that for both men and women, higher levels of mood instability predicted higher levels of self-reported risky cyber security behavior. For the remaining three of the five personality traits, none were related to self-reported cybersecurity knowledge or self-reported risky cybersecurity behaviors. In addition, we found that compared to women, men reported having higher levels of knowledge of secure passwords, $t(287) = -2.02$, $p = 0.04$, and $\eta^2 = 0.09$; lower levels of conscientiousness, $t(288) = 3.20$, $p = 0.002$, and $\eta^2 = 0.22$; lower levels of extraversion, $t(288) = 2.02$, $p = 0.04$, and $\eta^2 = 0.22$; lower levels of agreeableness, $t(288) = 3.33$, $p = 0.001$, and $\eta^2 = 0.20$; and higher levels of sensation-seeking personality, $t(289) = -5.08$, $p < 0.001$, and $\eta^2 = 0.17$.

To investigate further how self-reported knowledge about strong/weak passwords, personality traits, general risk-taking, and predict risky cybersecurity behaviors, we carried out a hierarchical multiple regression using risky cybersecurity behaviors as the dependent variable and four blocks of variables. Variables were examined to confirm that assumptions of normality, linearity, and homoscedasticity were met (Hair et al., 1998). We ordered the variables with a developmental trajectory of the individual in mind with personal characteristics entered in early blocks and self-reported knowledge and self-reported cybersecurity-related behaviors, in later blocks. This enabled us to examine the results for knowledge and cybersecurity behavior while controlling for personal characteristics and to examine cybersecurity behavior, while controlling for knowledge (Keith, 2014). In block one, sex was entered to control for sex differences. Subsequent blocks involved personality variables before knowledge, as both Big Five personality traits and sensation-seeking personality are generally believed to develop early in life and have a basis in biology (Fulker et al., 1980; Jang et al., 1996, respectively) and knowledge about technology acquired later. In block two, Big Five personality traits were added. In block three, sensation-seeking personality traits and

TABLE 1 | Summary of descriptive statistics and correlations for men (lower half of matrix) and women (upper half of matrix).

Variable	1	2	3	4	5	6	7	8	9	Mean	SE
1. Risky cybersecurity behavior	–	**−0.26*** **	−0.01	**−0.27*** **	−0.14	−0.10	**0.16* **	**0.49*** **	**0.15* **	2.92	0.09
2. Secure password knowledge	**−0.25* **	–	0.04	0.08	−0.06	0.10	0.08	−0.05	0.01	3.79	0.09
3. Extraversion	−0.01	−0.03	–	0.03	**0.21** **	0.10	0.02	**0.32*** **	**0.25*** **	5.88	0.10
4. Conscientiousness	−0.05	−0.01	**0.36*** **	–	**0.41*** **	**0.24*** **	**−0.27*** **	**−0.26*** **	**−0.20** **	6.44	0.08
5. Agreeableness	−0.03	0.13	0.19	**0.24* **	–	**0.32*** **	**−0.39*** **	**−0.18* **	−0.13	7.04	0.09
6. Openness	0.07	0.04	**0.22* **	**0.27** **	**0.34*** **	–	-0.01	0.07	**0.23** **	6.19	0.08
7. Mood instability	**0.21* **	−0.14	0.07	0.06	**−0.22* **	0.03	–	0.05	0.04	4.51	0.09
8. General risk-taking (DOSPERT)	**0.47*** **	0.12	0.19	−0.10	−0.12	0.10	+ 0.03	–	**0.50*** **	2.75	0.06
9. Sensation seeking personality	−0.10	0.12	0.19	0.09	−0.13	0.00	0.14	**0.20* **	–	16.18	0.45
Mean	2.86	4.10	5.52	5.99	6.58	6.40	4.31	3.32	19.72		
SE	0.10	0.12	0.14	0.11	0.11	0.11	0.10	0.08	0.49		

*Lower half of the matrix provides results for men and upper half provides results for women. *p < 0.05, **p < 0.01, ***p < 0.001. Bolded correlations are statistically significant.*

general risk-taking were added, and in block four, knowledge of password security was added. We found that variables did not involve excessive collinearity (as evidenced by the Tolerance and VIF values). Excessive collinearity would weaken the statistical power of the analysis (Coakes, 2005). **Table 2** displays the summary of these results.

In Block 1, participant sex did not significantly contribute to the variance in risky cybersecurity behaviors, $F(1, 287) = 0.01$ and $p = 0.91$. In Block 2, Big Five personality traits contributed significantly to variance in risky cybersecurity behaviors, accounting for 6% of the variance, $F(6, 287) = 3.11$ and $p = 0.006$ and the change in R^2 was significant, $F(5, 287) = 3.73$ and $p = 0.006$. Two of the five traits were significant predictors: (a) conscientiousness (β = −0.19 and $p = 0.03$) and (b) mood instability (β = 0.16 and $p = 0.02$). In Block 3, sensation-seeking personality traits and general risk-taking accounted for an additional 28% of variance in risky cybersecurity behaviors, $F(8, 287) = 13.70$ and $p < 0.001$, and the change in R^2 was significant, $F(2, 287) = 47.30$ and $p < 0.001$. Both variables were significant predictors: (a) sensation-seeking personality traits (β = −0.14 and $p = 0.02$) and (b) general risk-taking (β = 0.59 and $p < 0.001$). In Block 4, knowledge about secure passwords accounted for an additional 6% of variance in risky cybersecurity behaviors, $F(9, 287) = 17.48$ and $p < 0.001$, and the change in R^2 was significant, $F(1, 287) = 26.86$ and $p < 0.001$. Knowledge about security passwords was a significant predictor (β = −0.25 and $p < 0.001$). The total amount of variance accounted was 34%.

TABLE 2 | Summary of hierarchical regression analysis for variables predicting lax cybersecurity behaviors.

Variable	β	*T*	sr^2	*R*	*R*^{2}	Δ*R*^{2}
Block 1				0.004	0.00	0.00
Sex	−0.004	−0.07	0.00			
Block 2				0.25	0.04	0.06**
Sex	−0.02	−0.36	0.00			
Conscientiousness	−0.19	−2.94**	0.03			
Extraversion	0.01	0.10	0.00			
Agreeableness	0.04	0.56	0.01			
Openness	−0.02	−0.27	0.00			
Mood instability	0.16	2.58*	0.02			
Block 3				0.54	0.28	0.23***
Sex	−0.15	−2.56*	0.02			
Conscientiousness	−0.08	−1.30	0.00			
Extraversion	−0.14	−2.46*	0.02			
Agreeableness	0.13	2.02	0.01			
Openness	−0.08	−1.38	0.01			
Mood instability	0.17	3.16**	0.03			
Sensation-seeking traits	−0.13	−2.18*	0.01			
General risk-taking (DOSPERT)	0.59	9.54***	0.24			
Block 4				0.60	0.34	0.06***
Sex	−0.12	−2.19*	0.01			
Conscientiousness	−0.06	−1.10	0.00			
Extraversion	−0.14	−2.58*	0.02			
Agreeableness	0.12	2.03*	0.01			
Openness	−0.06	−1.12	0.00			
Mood instability	0.17	3.28***	0.03			
Sensation-seeking traits	−0.12	−2.12*	0.01			
General risk-taking (DOSPERT)	0.59	10.05***	0.24			
Password knowledge	−0.26	−5.23***	0.06			

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

DISCUSSION

The present research investigated how well self-reported risky cybersecurity behavior could be predicted by a combination of self-reported knowledge about secure passwords and personal characteristics, such as personality traits and general risk-taking in daily life. The majority of hypotheses tested were supported, including that (a) higher levels of self-reported password security knowledge was related to lower levels of self-reported risky cybersecurity behaviors, (b) higher levels of conscientiousness was related to lower levels of self-reported risky cybersecurity behaviors, (c) higher levels of mood instability was related to higher levels of self-reported risky cybersecurity behavior, (d) higher levels of sensation-seeking was related to higher levels of self-reported risky cybersecurity behaviors, (e) higher levels of general risk-taking behaviors was related to higher levels of risky cybersecurity behaviors, and (f) men reported engaging in more risk-taking in daily life than women, but the level of self-reported risky cybersecurity behavior did not differ for men and women. There were significant results that were not predicted. These include that for both men and women, higher levels of mood instability predicted higher levels of self-reported risky cyber security behavior; men reported having higher levels of password security knowledge than women.

The results showed that higher levels of self-reported password security knowledge was related to lower levels of self-reported risky cybersecurity behaviors, as has also been observed in prior research (Ferguson, 2005; McCrohan et al., 2010; Peker et al., 2016). Second, we found that women's higher levels of sensation-seeking, but not men's, were related to higher levels of self-reported risky cybersecurity behaviors for women. In prior research, sensation-seeking was not found to be related to cybersecurity behaviors (Whitty et al., 2015). Third, we found that higher levels of general risk-taking behaviors were related to higher levels of self-reported risky cybersecurity behaviors. Fourth, we found that conscientiousness predicted self-reported risky cybersecurity behaviors for women, but not men (cf. McCormac et al., 2017; Russell et al., 2017; Alohali et al., 2018; Shappie et al., 2019). The results also showed that higher levels of mood instability predicted higher levels of self-report risky cyber security behaviors, as has been observed in prior research (McCormac et al., 2017). We did not observe significant relationships between other three Big Five factors and risky cybersecurity behaviors. Our results showing that higher levels of sensation-seeking personality traits and general risk-taking in daily life predict greater use of risky cybersecurity behaviors are novel. These variables together contributed approximately 28% of the variance in cybersecurity behaviors, respectively. Overall, in our study in which 325 participants self-reported information about their password security knowledge, personality, risk-taking in daily life, and risky cybersecurity behavior, we found that personality variables and knowledge together predicted 34% of the variance in risky cybersecurity behaviors which exceeds the variance accounted for in prior research, which ranged between 3 and 5%.

The research has multiple strengths. It is the first to show that there is advantage to using personality traits in combination with other personal characteristics in predicting self-reported cybersecurity behavior. The statistical analysis provides estimates for the contributions of each. In addition, the present study is the first to show that general risk-taking in daily life predicts self-reported cybersecurity behavior. These results have implications for approaches in cybersecurity that involve training of individuals. These results suggest that creating profiles of potential victims of cybersecurity breaches should include personality variables, such as the Big Five and sensation-seeking, general risk-taking in daily life that is unrelated to using technology, in addition to knowledge about best practices in cybersecurity. The present results are the first to document that those who engage in higher levels of general risk-taking in daily life are also more likely to engage in risky cybersecurity behaviors. Our results suggest that accurate victim profiles could be useful in identifying individuals who are likely to be engaging in the highest levels of insecure cybersecurity behaviors. Institutions could use victim profiles to target such individuals with cybersecurity training that is in addition to what is typically taken. Other support from the institution could be targeted to those individuals. This approach is consistent with the view of Adams and Sasse (1999) who found that lack of cybersecurity knowledge and perceptions of risky behaviors as low risk could be viewed as the result of inadequate communication from institutional representatives to the users that they oversee. Institutions with high numbers of users with such victim profiles are encouraged to examine their communications to determine if improvements in communication can result in a reduction in numbers of users who fit victim profiles. Future research is needed to determine whether efforts to target individuals at high risk of being a cybersecurity victim with training or other support is effective in reducing their risk.

The present study also yielded some differences between men and women. As in prior research, men reported higher levels of sensation-seeking than women (Kennison et al., 2016) and higher levels of general risk-taking in daily life (Kennison et al., 2016). Nevertheless, we did not find that the level of self-reported risky cybersecurity behavior differed significantly. One prior study found that women engaged in risky cybersecurity behavior significantly less often than men (Anwar et al., 2017). Men reported significantly higher levels of password security knowledge. Prior research in which participants were drawn from employment settings have not observed differences in security knowledge, attitude, and behavior (McCormac et al., 2017). Prior research carried out on the online platform Amazon's Mechanical Turk (MTurk) also found that men reported significantly higher levels of knowledge than women (Cain et al., 2018).

There are multiple weaknesses of the research, including the characteristics of our sample. Our sample was majority female (i.e., 63.9%), relatively young (19.46 years on average), and drawn from university students enrolled in psychology and speech communication courses, which typically enroll more women than men. Our participants may be less aware of cybersecurity issues than others who are older or who are drawn from other settings. For this reason, the results may not generalize to other populations. A second limitation is that our assessments of risky cybersecurity behaviors and secure password knowledge were created for this study, and although the items for each construct demonstrated high internal consistency, they may fail to capture all aspects of risky cybersecurity behavior and/or secure password knowledge. The questions that we used to assess knowledge may have tapped into overlapping topics and may have reflected participants' opinion about their knowledge rather than actual knowledge. A third limitation is that we measured self-reported knowledge and cybersecurity behavior from participants. We may have observed different results had we been able to assess participants' knowledge and behavior using different methods. Future research is needed to determine whether our results are replicated in other samples and/or other populations. A fourth limitation may be the fact that the research was carried out in an online survey. It is possible that different results may be obtained when face-to-face survey methodologies are used.

Future research on this topic may improve on the present research in a number of ways. In the present study, we developed six-item questionnaire to assess behavior in situations familiar to college students and four-item questionnaire to assess participants' opinion about their cybersecurity knowledge. Future research could improve on the present research by assessing cybersecurity knowledge and behavior using objective measures instead of or in addition to self-report measures. Participants' responses to our questions about knowledge and behavior may not accurately measure either construct, but reflect a mixture of each construct and opinion, which may have led to participants responding in ways that they perceived to be socially more desirable. Future research could include measures to assess social desirability responding (e.g., Crowne and Marlowe, 1960). Future research could also include a wider variety of question types, such as open-ended questions that enable the researcher to assess participants' prior experiences (i.e., good or bad) with cybersecurity as well as other topics. The present research did not include open-ended questions about participants' past cybersecurity experiences.

In summary, the research showed that taking into consideration sensation-seeking personality traits and general risk-taking in daily life, in addition to participant sex, Big Five personality traits, and knowledge about security passwords accounts for about 34% of variance in risky cyber security behaviors. From previous work, this is one of the highest amounts of variance accounted for in cyber security behaviors. This greatly reinforces that more research is needed on the relationship between personality traits (or other traits) and cyber security behaviors. Institutions who rely on training to increase awareness about cybersecurity issues as a means to reduce risky cybersecurity behaviors may find that using personal characteristics to target training to individuals who are the most likely to engage in risky behaviors may lead to better return on investment. Some individuals are more likely to engage in risky cybersecurity behaviors than others. Better personalized cybersecurity training is needed from organizations to improve the cybersecurity compliance and cybersecurity behaviors of individuals.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Oklahoma State University Institutional Review Board (IRB). Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

SK and EC-T formulated the idea for the study, constructed items for the survey, and contributed to the writing of the manuscript. SK conducted the statistical analyses. Both authors contributed to the article and approved the submitted version.

FUNDING

This research was partially funded by a research grant to both authors from the National Science Foundation (DGE 1918591 and DGE 1919004).

REFERENCES

Adams, A., and Sasse, M. A. (1999). Users are not the enemy. *Commun. ACM* 42, 40–46. doi: 10.1145/322796.322806

Alohali, M., Clarke, N., Li, F., and Furnell, S. (2018). Identifying and predicting the factors affecting end-users' risk-taking behavior. *Inform. Comput. Secur.* 26, 306–326. doi: 10.1108/ICS-03-2018-0037

Anderson, G., and Brown, R. I. (1984). Real and laboratory gambling, sensation seeking and arousal. *Br. J. Psychol.* 75, 401–410. doi: 10.1111/j.2044-8295.1984.tb01910.x

Anwar, M., He, W., Ash, I., Yuan, X., Li, L., and Xu, L. (2017). Gender difference and employees' cybersecurity behaviors. *Comput. Hum. Behav.* 69, 437–443. doi: 10.1016/j.chb.2016.12.040

Ayyagari, R., and Tyks, J. (2012). Disaster at a university: a case study in information security. *J. Inform. Technol. Educ.* 11, 85–96. doi: 10.28945/1569

Bada, M., Sasse, A. M., and Nurse, J. R. (2019). *Cyber Security Awareness Campaigns: Why Do They Fail to Change Behaviour? arXiv* [Preprint]. Available online at: https://arxiv.org/abs/1901.02672 (accessed October 11, 2020).

Bailey, R. D., Foote, W. E., and Throckmorton, B. (2000). "Human sexual behavior: a comparison of college and Internet surveys," in *Psychological Experiments on the Internet*, ed. M. H. Birnbaum's (Cambridge, MA: Academic Press), 141–168.

Blais, A., and Weber, E. U. (2006). A domain-specific risk-taking (DOSPERT) scale for adult populations. *Judgm. Dec. Mak.* 1, 33–47. doi: 10.13072/midss.657

Blais, A. R., and Weber, E. U. (2001). Domain specificity and gender differences in decision making. *Risk Dec. Policy* 6, 47–69. doi: 10.1017/S1357530901000254

Bryant, K., and Campbell, J. (2006). User behaviours associated with password security and management. *Austr. J. Inform. Syst.* 14. Available online at: https://journal.acs.org.au/index.php/ajis/article/view/9

Buchanan, T., and Smith, J. L. (1999). Using the Internet for psychological research: personality testing on the World Wide Web. *Br. J. Psychol.* 90, 125–144. doi: 10.1348/000712699161189

Cain, A. A., Edwards, M. E., and Still, J. D. (2018). An exploratory study of cyber hygiene behaviors and knowledge. *J. Inform. Secur. Appl.* 42, 36–45. doi: 10.1016/j.jisa.2018.08.002

Coakes, S. J. (2005). *SPSS: Analysis Without Anguish*, 12 Edn, Hoboken, NJ: John Wiley & Sons.

Conley, J. J. (1985). Longitudinal stability of personality traits: a multi-trait-multimethod-multi-occasion analysis. *J. Person. Soc. Psychol.* 49, 1266–1282. doi: 10.1037/0022-3514.49.5.1266

Crowne, D. P., and Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *J. Consult. Psychol.* 24, 349–354. doi: 10.1037/h0047358

Dodou, D., and de Winter, J. C. (2014). Social desirability is the same in offline, online, and paper surveys: a meta-analysis. *Comput. Hum. Behav.* 36, 487–495. doi: 10.1016/j.chb.2014.04.005

Dwight, S. A., Cummings, K. M., and Glenar, J. L. (1998). Comparison of criterion-related validity coefficients for the Mini-Markers and Goldberg's markers of the big five Personality Factors. *J. Pers. Assess.* 70, 541–550. doi: 10.1207/s15327752jpa7003_11

Farcasin, M., and Chan-Tin, E. (2015). Why we hate IT: two surveys on pre-generated and expiring passwords in an academic setting. *Wiley Secur. Commun. Netw.* 8, 2361–2373. doi: 10.1002/sec.1184

Ferguson, A. J. (2005). Fostering e-mail security awareness: the west point carronade. *Educ. Q.* 28, 54–57.

Figner, B., and Weber, E. U. (2011). Who takes risks when and why? Determinants of risk taking. *Curr. Direct. Psychol. Sci.* 20, 211–216. doi: 10.1177/0963721411415790

Florencio, D., and Herley, C. (2007). "A large-scale study of web password habits," in *Proceedings of the 16th international conference on World Wide Web*, New York, NY.

Frey, R., Pedroni, A., Mata, R., Rieskamp, J., and Hertwig, R. (2017). Risk preference shares the psychometric structure of major psychological traits. *Sci. Adv.* 3:e1701381. doi: 10.1126/sciadv.1701381

Fulker, D. W., Eysenck, S. B., and Zuckerman, M. (1980). A genetic and environmental analysis of sensation seeking. *J. Res. Pers.* 14, 261–281. doi: 10.1016/0092-6566(80)90033-1

Gaw, S., and Felten, E. W. (2006). "Password management strategies for online accounts," in *Proceedings of the Second Symposium on Usable Privacy and Security*, New York, NY.

George, D., and Mallery, P. (2003). *SPSS for Windows Step by Step: A Simple Guide and Reference. 11.0 Update*, 4th Edn, Boston, MA: Allyn & Bacon.

Gosling, S., Vazire, S., Srivastava, S., and John, O. (2004). Should we trust web-based studies? A comparative analysis of six preconceptions about internet questionnaires. *Am. Psychol.* 59, 93–104. doi: 10.1037/0003-066x.59.2.93

Grawemeyer, B., and Johnson, H. (2011). Using and managing multiple passwords: a week to a view. *Interact. Comput.* 23, 256–267. doi: 10.1016/j.intcom.2011.03.007

Gurven, M., Von Rueden, C., Massenkoff, M., Kaplan, H., and Lero Vie, M. (2013). How universal is the Big Five? Testing the five-factor model of personality variation among forager-farmers in the Bolivian Amazon. *J. Pers. Soc. Psychol.* 104:354. doi: 10.1037/a0030841

Gustafsod, P. E. (1998). Gender Differences in risk perception: theoretical and methodological perspectives. *Risk Analys.* 18, 805–811. doi: 10.1023/b:rian.0000005926.03250.c0

Hair, J. F., Anderson, R. E., Tatham, R. L., and Black, W. C. (1998). *Multivariate Data Analysis*, 5th Edn, Upper Saddle River, NJ: Prentice Hall.

Horvath, P., and Zuckerman, M. (1993). Sensation seeking, risk appraisal, and risky behavior. *Pers. Individ. Differ.* 14, 41–52. doi: 10.1016/0191-8869(93)90173-Z

Jang, K. L., Livesley, W. J., and Vemon, P. A. (1996). Heritability of the big five personality dimensions and their facets: a twin study. *J. Pers.* 64, 577–592. doi: 10.1111/j.1467-6494.1996.tb00522.x

Keith, T. Z. (2014). *Multiple Regression and Beyond: An Introduction to Multiple Regression and structUral Equation Modeling*. Abingdon: Routledge.

Kennison, S. M., and Messer, R. H. (2017). Cursing as a form of risk-taking. *Curr. Psychol.* 36, 119–126. doi: 10.1007/s12144-015-9391-1

Kennison, S. M., and Messer, R. H. (2019). Humor as social risk-taking: the relationships among humor styles, sensation-seeking, and use of curse words. *Humor* 32, 1–21. doi: 10.1515/humor-2017-0032

Kennison, S. M., Wood, E. E., Byrd-Craven, J., and Downing, M. L. (2016). Financial and ethical risk-taking by young adults: a role for family dynamics during childhood. *Cogent Econ. Finan.* 4:1232225. doi: 10.1080/23322039.2016.1232225

Kraut, R., Olson, J., Banaji, M., Bruckman, A., Cohen, J., and Couper, M. (2004). Psychological research online: report of board of scientific affairs' advisory group on the conduct of research on the internet. *Am. Psychol.* 59, 105–117. doi: 10.1037/0003-066X.59.2.105

Lorenz, B., Kikkas, K., and Klooster, A. (2013). "The four most-used passwords are love, sex, secret, and god: Password security and training in different user groups," in *Proceedings of the International Conference on Human Aspects of Information Security, Privacy, and Trust*, Cham.

Maraj, A., Martin, M. V., Shane, M., and Mannan, M. (2019). "On the null relationship between personality types and passwords," in *Proceedings of the 7th International Conference on Privacy, Security and Trust (PST)*, Fredericton, NB.

McBride, M., Carter, L., and Warkinten, M. (2012). *Exploring the Role of INDIVIDUAL employee Characteristics and Personality on Employee Compliance with Cyber Security Policies*. Triangle Park, CA: RTI International-Institute for Homeland Security Solutions.

McCormac, A., Zwaans, T., Parsons, K., Calic, D., Butavicius, M., and Pattinson, M. (2017). Individual differences and information security awareness. *Comput. Hum. Behav.* 69, 151–156. doi: 10.1016/j.chb.2016.11.065

McCrohan, K. F., Engel, K., and Harvey, J. W. (2010). Influence of awareness and training on cyber security. *J. Internet Commer.* 9, 23–41. doi: 10.1080/15332861.2010.487415

Mitnick, K. D. (2003). Are you the weak link?. *Harvard Bus. Rev.* 81, 18–20.

Mooradian, T. A., and Nezlek, J. B. (1996). Comparing the NEO-FFI and Saucier's Mini-Markers as measures of the Big Five. *Pers. Individ. Differ.* 21, 213–215. doi: 10.1016/0191-8869(96)00057-8

Notoatmodjo, G., and Thomborson, C. (2009). "Passwords and perceptions," in *Proceedings of the Seventh Australasian Conference on Information Security*, Wellington.

Nunnally, J. C. (1978). *Psychometric Theory*, 2nd Edn, New York, NJ: McGraw-Hill.

Panno, A., Donati, M. A., Milioni, M., Chiesi, F., and Primi, C. (2018). Why women take fewer risk than men do: the mediating role of state anxiety. *Sex Roles* 78, 286–294. doi: 10.1007/s11199-017-0781-8

Peker, Y. K., Ray, L., Da Silva, S., Gibson, N., and Lamberson, C. (2016). Raising cybersecurity awareness among college students. *J. Colloq. Inform. Syst. Secur. Educ.* 4, 1–17. doi: 10.1201/9780429031908-1

Pew Research Center (2017). *Americans and Cybersecurity*. Available online at: https://www.pewresearch.org/internet/2017/01/26/americans-and-cybersecurity/ (accessed October 11, 2020).

Plachkinova, M., and Maurer, C. (2019). Security breach at target. *J. Inform. Syst. Educ.* 29:7.

Popham, L., Kennison, S. M., and Bradley, K. I. (2011). Ageism, sensation-seeking, and risk-taking in young adults. *Curr. Psychol.* 30, 184–193. doi: 10.1007/s12144-001-9107-0

Proctor, W. R. (2016). *Investigating the Efficacy of Cybersecurity Awareness Training Programs*. Doctoral thesis, Utica College, Utica, NY.

Ramlo, S. E., and Nicholas, J. B. (2020). Divergent student views of cybersecurity. *J. Cybersecur. Educ. Res. Pract.* 2019:6.

Riley, S. (2006). Password security: what users know and what they actually do. *Usabil. News* 8, 2833–2836.

Roberti, J. W. (2004). A review of behavioral and biological correlates of sensation seeking. *J. Res. Pers.* 38, 256–279. doi: 10.1016/S0092-6566(03)00067-9

Russell, J. D., Weems, C. F., Ahmed, I., and Richard, G. G. III (2017). Self-reported secure and insecure cyber behaviour: factor structure and associations with personality factors. *J. Cyber Secur. Technol.* 1, 163–174. doi: 10.1080/23742917.2017.1345271

Saucier, G. (1994). Mini-Markers: a brief version of Goldberg's unipolar big-five markers. *J. Pers. Assess.* 63, 506–516. doi: 10.1207/s15327752jpa6303_8

Shappie, A. T., Dawson, C. A., and Debb, S. M. (2019). Personality as a predictor of cybersecurity behavior. *Psychol. Pop. Med. Cult.* 9, 475–480. doi: 10.1037/ppm0000247

Shou, Y., and Olney, J. (2020). Assessing a domain-specific risk-taking construct: a meta-analysis of reliability of the DOSPERT scale. *Judg. Dec. Mak.* 15:112.

Stobert, E., and Biddle, R. (2014). "The password life cycle: user behaviour in managing passwords," in *Proceedings of the 10th Symposium on Usable Privacy and Security (SOUPS 2014)*, Cham.

Taber, K. S. (2018). The use of cronbach's alpha when developing and reporting research instruments in science education. *Res. Sci. Educ.* 48, 1273–1296. doi: 10.1007/s11165-016-9602-2

Tamrakar, A., Russell, J. D., Ahmed, I., Richard, G. G. III, and Weems, C. F. (2016). "SPICE: A software tool for bridging the gap between end-user's insecure cyber behavior and personality traits," in *Proceedings of the Sixth ACM Conference on Data and Application Security and Privacy*, New York, NY.

Taylor-Jackson, J., McAlaney, J., Foster, J., Bello, A., Maurushat, A., and Dale, J. (2020). "Incorporating psychology into cyber security education: a pedagogical approach," in *Proceedings of Asia USEC'20, Financial Cryptography and Data Security*, Sabah.

Thomas, K., Li, F., Zand, A., Barrett, J., Ranieri, J., Invernizzi, L., et al. (2017). "Data breaches, phishing, or malware? Understanding the risks of stolen credentials," in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, Hoboken, NJ.

Wang, P., and Johnson, C. (2018). Cybersecurity incident handling: a case study of the Equifax data breach. *Issues Inform. Syst.* 19, 150–159.

Weber, E. U., Blais, A.-R., and Betz, E. (2002). A domain specific risk-attitude scale: measuring risk perceptions and risk behaviors. *J. Behav. Dec. Mak.* 15, 263–290. doi: 10.1002/bdm.414

Weigold, A., Weigold, I. K., and Russell, E. J. (2013). Examination of the equivalence of self-report survey-based paper-and-pencil and internet data collection methods. *Psychol. Methods* 18, 53–70. doi: 10.1037/a0031607

Whitty, M., Doodson, J., Creese, S., and Hodges, D. (2015). Individual differences in cyber security behaviors: an examination of who is sharing passwords. *Cyberpsychol. Behav. Soc. Netw.* 18, 3–7. doi: 10.1089/cyber.2014.0179

Yamagata, S., Suzuki, A., Ando, J., Ono, Y., Kijima, N., Yoshimura, K., et al. (2006). Is the genetic structure of human personality universal? A cross-cultural twin study from North America, Europe, and Asia. *J. Pers. Soc. Psychol.* 90, 987–998. doi: 10.1037/0022-3514.90.6.987

Zuckerman, M. (1983a). Sensation seeking and sports. *Pers. Individ. Differ.* 4, 285–292. doi: 10.1016/0191-8869(83)90150-2

Zuckerman, M. (1983b). "Sensation seeking: the initial motive for drug abuse," in *Etiological Aspects of Alcohol and Drug Abuse*, eds E. Gotheil, K. A. Druley, T. E. Skoloda, and H. M. Waxman (New York, NY: Thomas), 202–220.

Zuckerman, M. (1987). "Is sensation seeking a predisposing trait for alcoholism?," in *Stress and Addiction*, eds E. Gottheil, K. A. Druley, S. Pashkey, and S. P. Weinstein (Levittown, PA: Mazel), 283–301.

Zuckerman, M. (1994). *Behavioural Expressions and Biosocial Bases of Sensation-Seeking*. Cambridge: Cambridge University Press.

Zuckerman, M., Ball, S., and Black, J. (1990). Influences of sensation seeking, gender, risk appraisal, and situational motivation on smoking. *Add. Behav.* 15, 209–220. doi: 10.1016/0306-4603(90)90064-5

Zuckerman, M., Eysenck, S., and Eysenck, H. J. (1978). Sensation seeking in England and America: cross-cultural, age, and sex comparisons. *J. Consult. Clin. Psychol.* 46, 139–149. doi: 10.1037/0022-006x.46.1.139

Zuckerman, M., Kolin, E. A., Price, L., and Zoob, I. (1964). Development of a sensation-seeking scale. *J. Consult. Psychol.* 28, 477–482. doi: 10.1037/h0040995

Zuckerman, M., and Link, K. (1968). Construct validity for the sensation-seeking scale. *J. Consult. Clin. Psychol.* 32:420. doi: 10.1037/h0026047

Zuckerman, M., Tushup, R., and Finner, S. (1976). Sexual attitudes and experience: attitude and personality correlations and changes produced by a course in sexuality. *J. Consult. Clin. Psychol.* 44, 7–19. doi: 10.1037/0022-006X.44.1.7

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 07 December 2020
doi: 10.3389/fpsyg.2020.528079

Repetition of Computer Security Warnings Results in Differential Repetition Suppression Effects as Revealed With Functional MRI

C. Brock Kirwan[1,2], Daniel K. Bjornn[2], Bonnie Brinton Anderson[3], Anthony Vance[4], David Eargle[5] and Jeffrey L. Jenkins[3]*

[1] Neuroscience Center, Brigham Young University, Provo, UT, United States, [2] Department of Psychology, Brigham Young University, Provo, UT, United States, [3] Department of Information Systems, Brigham Young University, Provo, UT, United States, [4] Department of Management Information Systems, Fox School of Business, Temple University, Philadelphia, PA, United States, [5] Leeds School of Business, University of Colorado Boulder, Boulder, CO, United States

OPEN ACCESS

Edited by:
David Maimon,
Georgia State University,
United States

Reviewed by:
Irina M. Harris,
The University of Sydney, Australia
Zarul Fitri Zaaba,
Universiti Sains Malaysia (USM),
Malaysia
Eric A. Walden,
Texas Tech University, United States
Adriane B. Randolph,
Kennesaw State University,
United States

***Correspondence:**
C. Brock Kirwan
kirwan@byu.edu

Specialty section:
This article was submitted to Cognition, a section of the journal Frontiers in Psychology

***Received:** 18 January 2020*
***Accepted:** 11 November 2020*
***Published:** 07 December 2020*

Citation:
Kirwan CB, Bjornn DK, Anderson BB, Vance A, Eargle D and Jenkins JL (2020) Repetition of Computer Security Warnings Results in Differential Repetition Suppression Effects as Revealed With Functional MRI.
Front. Psychol. 11:528079.
doi: 10.3389/fpsyg.2020.528079

Computer users are often the last line of defense in computer security. However, with repeated exposures to system messages and computer security warnings, neural and behavioral responses show evidence of habituation. Habituation has been demonstrated at a neural level as repetition suppression where responses are attenuated with subsequent repetitions. In the brain, repetition suppression to visual stimuli has been demonstrated in multiple cortical areas, including the occipital lobe and medial temporal lobe. Prior research into the repetition suppression effect has generally focused on a single repetition and has not examined the pattern of signal suppression with repeated exposures. We used complex, everyday stimuli, in the form of images of computer programs or security warning messages, to examine the repetition suppression effect across repeated exposures. The use of computer warnings as stimuli also allowed us to examine the activation of learned fearful stimuli. We observed widespread linear decreases in activation with repeated exposures, suggesting that repetition suppression continues after the first repetition. Further, we found greater activation for warning messages compared to neutral images in the anterior insula, pre-supplemental motor area, and inferior frontal gyrus, suggesting differential processing of security warning messages. However, the repetition suppression effect was similar in these regions for both warning messages and neutral images. Additionally, we observed an increase of activation in the default mode network with repeated exposures, suggestive of increased mind wandering with continuing habituation.

Keywords: repetition suppression, fMRI, habituation, anterior insula, cybersecurity

INTRODUCTION

One major obstacle to computer security is habituation on the part of computer users to repeated computer security messages. Sometimes termed "warning fatigue," this habituation to security warnings can result in lower rates of security behavior (Akhawe and Felt, 2013). At a biological level, repeated exposure to a stimulus results in repetition suppression, or a decreased neuronal response to that stimulus. Evidence for repetition suppression has been observed for both auditory (Costa-Faidella et al., 2011; Todorovic et al., 2011) and visual processing (Summerfield et al., 2008, 2011; Larsson and Smith, 2012) using recording methods including single-unit recording electrophysiology (Malmierca et al., 2009), functional magnetic resonance imaging (fMRI) (Larsson and Smith, 2012; Grotheer and Gyula, 2015), electroencephalography (Costa-Faidella et al., 2011; Summerfield et al., 2011), and magnetoencephalography (Todorovic et al., 2011; Todorovic and de Lange, 2012).

The effect of habituation has been studied in different ways in different fields. For example, in marketing, a great deal of research has studied "repetition effects" (Schmidt and Eisend, 2015), or the "differential effects of each successive advertising exposure, i.e., the differential effects of a given exposure within a sequence of exposures" (Pechmann and Stewart, 1988, p. 287). The most accepted theory explaining repetition effects is Berlyne's (1970) two-factor theory that explains a "wear-in" process in which familiarity and ad effectiveness increases with repetitions, and a later "wear-out" process, in which the effectiveness of an advertisement decreases with each succeeding exposure.

In contrast, in the fields of warning science and computer security, repeated exposure to a warning does not lead to beneficial familiarity effects, but leads directly to diminished attention to a warning (Wogalter and Vigilante, 2006; Vance et al., 2017). In computer security, habituation to warnings has been frequently inferred as a factor without measuring it directly (Bravo-Lillo et al., 2014). For example, Akhawe and Felt (2013, p. 268) reported that the most common web browser SSL error had the lowest adherence rate, which they concluded was "indicative of warning fatigue." However, some studies have examined habituation directly by measuring decreased attention to warnings using eye-tracking, mouse cursor tracking, and fMRI (Anderson et al., 2016a,b; Vance et al., 2018). The results from all of these studies show decrease attention to warnings after only 2–3 exposure. However, none of these studies directly compared how people habituate to computer security warning stimuli compared to general software application stimuli, a gap that this article investigates.

The underlying process of repetition suppression is not fully known and there is some debate as to the mechanisms that achieve the decrease in neuronal activation. One view is the bottom-up, or fatigue model, which suggests that differences in activity are related to the refractory period of local neural generators in response to physical stimulation (see Grill-Spector et al., 2006, for review). Another view is the top-down, or predictive coding, model which posits that repetition suppression is due to the expected probability of a stimulus recurring (Mayrhauser et al., 2014). Recent research gives support for the predictive coding model; Summerfield et al. (2008) found that the repetition suppression effect was modulated by an expectation of how often stimuli would repeat. Larsson and Smith (2012) also found that expectation can influence the repetition suppression effect, but only when one is actively attending to the repeated stimulus. Valentini (2011), however, observes that there is evidence for some contribution by both bottom-up and top-down processes in repetition suppression.

The response to repeated stimulus exposure is not uniform across the brain and may depend on context or task demands. Multiple areas in the occipital and temporal lobes demonstrate a repetition suppression effect (Kovacs et al., 2013; Mayrhauser et al., 2014). Structures in the medial temporal lobe (MTL) including the hippocampus also demonstrate decreased fMRI activation in response to repeated stimuli, sometimes referred to as a novelty response (Stern et al., 1996). On the other hand, other regions of the MTL demonstrate an increase in fMRI activation in response to repeated stimuli (Kirwan et al., 2009), referred to as a familiarity response (e.g., Daselaar et al., 2006). In a review of the repetition enhancement effect (increased fMRI activation with stimulus repetition), Segaert et al. (2013) identified several factors that influence whether repetition suppression or repetition enhancement is observed. These factors include task demands and cognitive processes engaged (including memory, learning, and attention). Further, regions in the default mode network (DMN), including the medial parietal lobe, inferior parietal lobule, and prefrontal cortex, also demonstrate an increase in fMRI activation with repeated stimulus exposure (Danckert et al., 2007; McDonald et al., 2010). This increase in DMN activation has been linked to inattention to a specific stimulus (Mason et al., 2007; Raichle and Snyder, 2007) as demonstrated by decreased subsequent recognition memory accuracy (Shrager et al., 2008). Based on these findings, it is reasonable to assume that repeated exposure to a stimulus will result in decreased activation in sensory and attention networks and increased activation in the DMN.

Studies of repetition suppression typically use only a few repetitions over a short period of time typically lasting only a few minutes (Chouinard et al., 2008; Summerfield et al., 2008, 2011). Further, while some studies of novelty and familiarity effects have demonstrated both effects in different regions of the MTL within the same paradigm (notably in the hippocampus; e.g., Daselaar et al., 2006), none have examined the longer-term trade-off between novelty and familiarity signaling in the same region within the same paradigm. Thus, it is unclear if these repetition suppression effects (i.e., decreases in fMRI activation) would continue with repeated exposures to the same stimulus in the same scanning session.

Another limitation of the current repetition suppression effect literature is that generally simple stimuli have been studied, such as tones (Costa-Faidella et al., 2011; Todorovic et al., 2011; Todorovic and de Lange, 2012) or single objects (Chouinard et al., 2008; Kovacs et al., 2013). More complex visual stimuli, such as faces, have been used as well (Summerfield et al., 2008, 2011; Larsson and Smith, 2012). However, it is not known how repetition applies to complex, everyday stimuli such as images of computer programs over repeated exposures, much like what is experienced during everyday computer use. Accordingly, images of common computer scenes provide a real-world application for the phenomenon of repetition suppression. Further, computer security warning messages have a learned, negative emotional content. Thus, the use of computer warning messages provides the opportunity to examine the effect of learned emotional stimuli in a more realistic setting.

Computer security warnings are not inherently aversive stimuli and thus any negative emotional valence associated with them must be learned, likely through social or verbal means. While much is known about the neural circuitry involved in classical fear conditioning, relatively little is known about the neural circuitry of social fear learning (Olsson and Phelps, 2007). Classical fear conditioning is critically dependent on the amygdala (Phillips and LeDoux, 1992) and has been shown to activate amygdala in human neuroimaging paradigms (Buchel and Dolan, 2000). Similarly, social and verbal fear learning have been shown to activate the amygdala (e.g., Phelps et al., 2001), indicating a general role of the amygdala in fear acquisition and fear expression in both classically conditioned and socially or verbally acquired fear responses. The anterior insula is also activated for verbally acquired fear representations (Phelps et al., 2001; Olsson and Phelps, 2007). Anterior insula activity has been linked to the anticipation of negative events (Grupe and Nitschke, 2013) and its dysfunction has been linked to avoidance of threat uncertainty (Paulus and Stein, 2006). Anterior insula activation has also been associated with general arousal levels, regardless of positive or negative valence of the stimulus (Knutson et al., 2014).

In the current experiment, we sought to examine the repetition suppression effect over repeated exposures to complex, everyday stimuli both generally and for socially constructed fearful stimuli. We anticipated a repetition suppression effect (i.e., decreased BOLD signal) in visual processing stream but increased activation in DMN regions with repeated exposures. We further examined the effect of repeated exposures on novelty and familiarity signals in the MTL. Finally, we investigated the effects of repetition on responses of brain regions associated with fear and/or arousal.

MATERIALS AND METHODS

Subjects

Twenty-two participants (4 female, 18 male; 24 years old, range 20–27) were recruited from the university community and gave written informed consent prior to participation. The sample size was determined by previous literature in this area (Dimoka, 2012) and guidelines set forth by Desmond and Glover (2002) to calculate the required number of subjects to ensure adequate statistical power. Participants were right-handed native English speakers with normal or corrected-normal visual acuity. Participants self-reported free of psychiatric or neurological conditions. As members of the university community, these subjects had a high level of computer literacy. The experiment was approved by the University Institutional Review Board and was conducted in accordance with the principles of the Declaration of Helsinki. Participants were compensated US $25 for a 60 min session.

Behavioral Task

We used an event-related, within-subject experimental design in which participants viewed a random sequence of 60 images of general software application screenshots (such as Microsoft Word, Excel, and other common applications) and security warnings collected by the researchers (**Figure 1**). The experiment utilized a variety of actual security warnings from programs running on a Windows operating system. **Table 1** summarizes each type of warning.

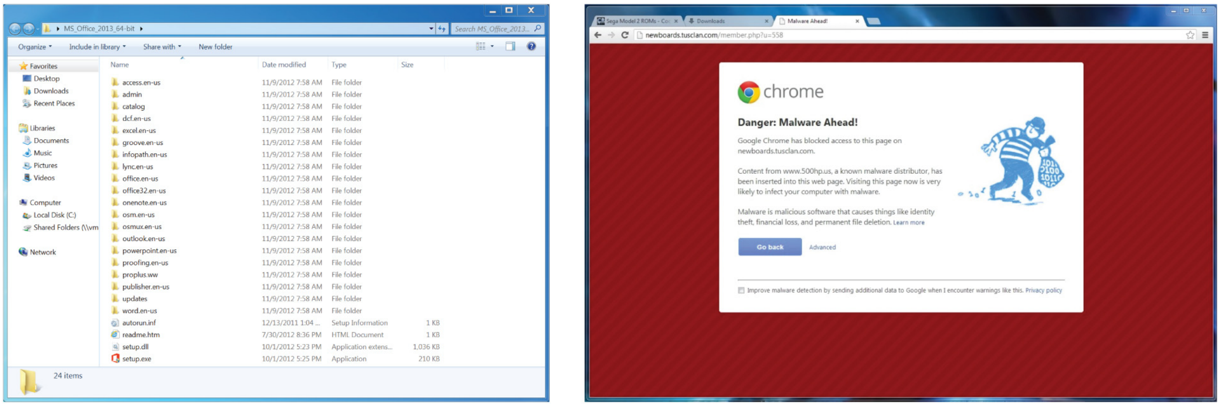
For visual consistency, all images of general software applications and security warnings were for the Windows operating system. Our experimental design is graphically depicted in **Figure 2** and consisted of two steps for each participant. In Step 1, images were organized into three sets of 20 images each. The first two sets comprised security warnings and general software applications. These were repeated six times each in random order across the duration of the scan. A third set consisted of general software application images, which were each displayed only once during the scan. This was done to create a baseline of unique presentations throughout the task. Thus, there were 260 total images (20 warnings × 6 repetitions + 20 software images × 6 repetitions + 20 software images × 1 exposure each) displayed in the experiment. In Step 2, the 260 images were randomized for each participant across two blocks of 7.7 min each (with a ∼2 min break in between).

Subjects were given a verbal briefing about the MRI procedures and the task, and then situated supine in the MRI scanner. Visual stimuli were displayed using E-prime software (version 2.0.10) and were viewed by means of a mirror attached to the head coil reflecting a large monitor outside the scanner. On each trial, images were displayed for 3 s each, with a 0.5 s inter-stimulus interval (ISI).

In order to keep participants attentive during the viewing of images, they were instructed to use an MR-compatible keypad to indicate if the image shown was common or uncommon in their experience. We intentionally used a simple task in order to minimize influence on the repetition suppression effect, while still enabling measurement of participant attention to the task. Such an approach is common in pattern separation tasks, for example, where the repetition suppression effect is used to differentiate repeated images from similar lures (Lacy et al., 2011). Participants responded on 96% of trials (*SD* = 10%), indicating that they were appropriately engaged and on task. At the end of the experimental task, participants were debriefed, compensated, and dismissed.

Equipment and Scan Parameters

MRI scanning took place on a Siemens 3T Trio scanner. For each scanned subject, we collected a high-resolution structural MRI scan for functional localization in addition to the two functional scans. Structural images were acquired with a T1-weighted magnetization-prepared rapid acquisition with gradient echo (MP-RAGE) sequence with the following parameters: TE = 2.26 ms, flip angle = 9°, slices = 176, slice thickness = 1.0 mm, matrix size = 256 × 215, voxel size = 1 mm × 0.98 mm × 0.98 mm. Functional scans were acquired with a gradient-echo, echo-planar, T2*-weighted pulse sequence with the following parameters: TR = 2,000 ms, TE = 28 ms, flip angle = 90°, slices = 40, slice thickness = 4.0 mm (no skip), matrix size = 64 × 64, voxel size = 3.44 mm × 3.44 mm × 3 mm. All data are available at https://openneuro.org/datasets/ds002363 and data analysis scripts are available at https://github.com/Kirwanlab/RepetitionSuppression.



Name	Date modified	Type	Size
access.en-us	11/9/2012 7:58 AM	File folder	
admin	11/9/2012 7:58 AM	File folder	
catalog	11/9/2012 7:58 AM	File folder	
dcf.en-us	11/9/2012 7:58 AM	File folder	
excel.en-us	11/9/2012 7:58 AM	File folder	
groove.en-us	11/9/2012 7:58 AM	File folder	
infopath.en-us	11/9/2012 7:58 AM	File folder	
lync.en-us	11/9/2012 7:58 AM	File folder	
office.en-us	11/9/2012 7:58 AM	File folder	
office32.en-us	11/9/2012 7:58 AM	File folder	
onenote.en-us	11/9/2012 7:58 AM	File folder	
osm.en-us	11/9/2012 7:58 AM	File folder	
osmux.en-us	11/9/2012 7:58 AM	File folder	
outlook.en-us	11/9/2012 7:58 AM	File folder	
powerpoint.en-us	11/9/2012 7:58 AM	File folder	
proofing.en-us	11/9/2012 7:58 AM	File folder	
proplus.ww	11/9/2012 7:58 AM	File folder	
publisher.en-us	11/9/2012 7:58 AM	File folder	
updates	11/9/2012 7:58 AM	File folder	
word.en-us	11/9/2012 7:58 AM	File folder	
autorun.inf	12/13/2011 1:04 ...	Setup Information	1 KB
readme.htm	7/30/2012 8:36 PM	HTML Document	1 KB
setup.dll	10/1/2012 5:23 PM	Application extens...	1,036 KB
setup.exe	10/1/2012 5:25 PM	Application	210 KB

24 items

newboards.tusclan.com/member.php?u=558

Danger: Malware Ahead!

Google Chrome has blocked access to this page on newboards.tusclan.com.

Content from www.500hp.us, a known malware distributor, has been inserted into this web page. Visiting this page now is very likely to infect your computer with malware.

Malware is malicious software that causes things like identity theft, financial loss, and permanent file deletion. Learn more

□ Improve malware detection by sending additional data to Google when I encounter warnings like this. Privacy policy

FIGURE 1 | Example stimuli from the behavioral experiment. Participants viewed repetitions of general software images (left) and computer security warnings (right) in a randomized order.

TABLE 1 | Description of warnings shown to participants.

Warning type description
The operating system warning that a program can "make changes to this computer"
A virus protection program warning "intruder detected"
A firewall warning "Danger: Malware Ahead!"
A firewall warning "blocked activity of harmful software"
Facebook warning of a potentially abusive link
A firewall warning that it has block some feature of a program
A web browser warning that a page contains non-secure items
A spreadsheet warning that a file contains macros
A web browser warning of a "Reported Web Forgery"
A program warning that an online application is attempting to access files on your computer
The operating system warning that an application is trying to run
The operating system warning that it cannot verify the publisher of a driver software
A browser warning that a connection is untrusted (SSL warning)
A virus protection program warning that a trojan was found

Analysis

MRI data were analyzed with the Analysis of Functional Images (AFNI) suite of programs (Cox, 1996). Briefly, structural and functional scans were converted to NIfTI file format using dcm2niix[1] (Li et al., 2016) which performs slice time correction of functional scans as part of the conversion process. Motion correction of the functional runs was calculated based on the volume with the least amount of noise for each functional run. Spatial normalization was calculated for each T1-weighted structural scan to MNI space. The motion correction and spatial normalization transformations were concatenated so that functional data underwent a single interpolation, thus reducing blurring of the data in preprocessing (Muncy et al., 2017). Functional data were scaled by the mean signal intensity. An intersection mask was calculated based on the overlap of the extent of coverage of the T2*-weighted functional scans and a gray matter mask of the MNI template brain. All group analyses were performed within this intersection mask.

For the first-level regression analysis, behavioral vectors were created that coded for stimulus type (e.g., security warnings, general software application screenshots) and repetition number. Additionally, we included a regressor for the single-presentation general computing screenshots to serve as a stimulus check to ensure that any observed decreases in responding were not due to fatigue. Stimulus events were modeled using a 3 s boxcar function convolved with the canonical hemodynamic response. Regressors coding for motion (6 regressors per scan run) and polynomial regressors coding for scan run and scanner drift were also entered into the model as nuisance variables. To control for size differences between the general software application screenshots and security warnings, the total size of each stimulus (in pixels) was also entered as a nuisance variable. Resulting beta values were blurred with a 5 mm FWHM Gaussian kernel. Beta values for the conditions of interest were then entered into the group-level analysis, which consisted of a model with stimulus type (two levels) and repetition number (six levels) as within-subject factors. The residuals from the first-level regression analysis were also blurred with a 5 mm FWHM Gaussian kernel and used to estimate the smoothness of each functional scan. This smoothness estimate was then entered into Monte Carlo simulations to determine a spatial extent threshold for performing corrections for multiple comparisons in group-level analyses (Cox et al., 2017). All tests were corrected for multiple comparisons using a voxel-wise threshold of $p < 0.001$ and a spatial extent threshold of 12 voxels, nearest-neighbors level 2 (overall $p < 0.01$).

RESULTS

As our hypotheses concerned differential responses over repeated exposures to stimuli, we first identified clusters that showed a main effect of repetition. Fifteen clusters survived correction for multiple comparison: left and right dorsal

[1] https://github.com/rordenlab/dcm2niix

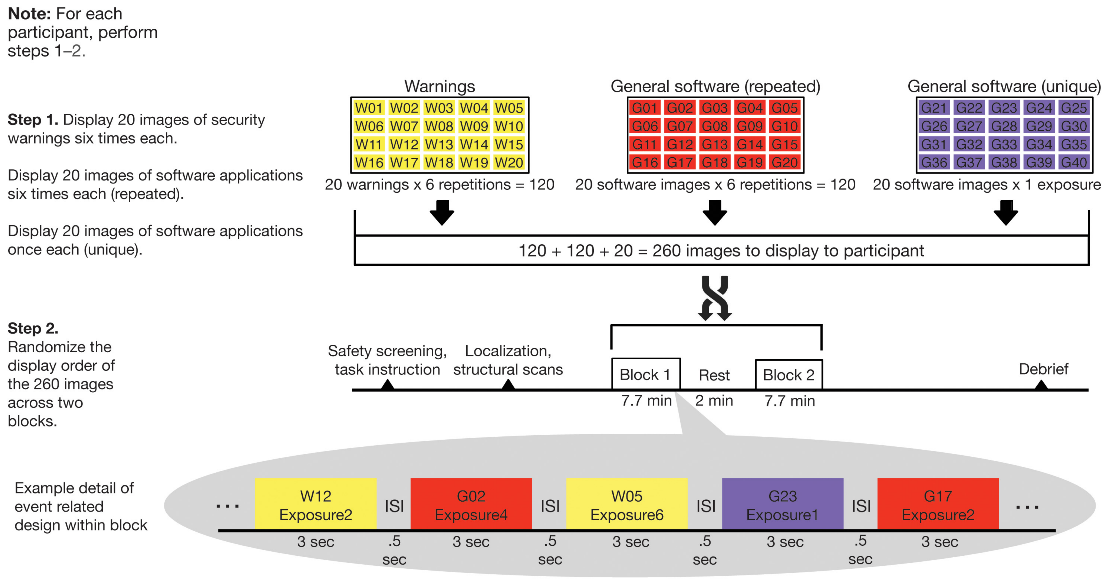

Note: For each participant, perform steps 1–2.
Step 1. Display 20 images of security warnings six times each.
Display 20 images of software applications six times each (repeated).
Display 20 images of software applications once each (unique).
Warnings
W01 W02 W03 W04 W05
W06 W07 W08 W09 W10
W11 W12 W13 W14 W15
W16 W17 W18 W19 W20
20 warnings x 6 repetitions = 120
General software (repeated)
G01 G02 G03 G04 G05
G06 G07 G08 G09 G10
G11 G12 G13 G14 G15
G16 G17 G18 G19 G20
20 software images x 6 repetitions = 120
General software (unique)
G21 G22 G23 G24 G25
G26 G27 G28 G29 G30
G31 G32 G33 G34 G35
G36 G37 G38 G39 G40
20 software images x 1 exposure = 20
120 + 120 + 20 = 260 images to display to participant
Step 2. Randomize the display order of the 260 images across two blocks.
Safety screening, task instruction
Localization, structural scans
Block 1
7.7 min
Rest
2 min
Block 2
7.7 min
Debrief
Example detail of event related design within block
...
W12 Exposure2
3 sec
ISI
.5 sec
G02 Exposure4
3 sec
ISI
.5 sec
W05 Exposure6
3 sec
ISI
.5 sec
G23 Exposure1
3 sec
ISI
.5 sec
G17 Exposure2
3 sec
...

FIGURE 2 | Randomization and behavioral task scheme. Computer security warning and general software images were displayed in two blocks of 130 images in random order. Twenty security warnings and 20 general software images were displayed six times each while 20 additional general software images were displayed once only.

and ventral visual processing streams, left and right inferior frontal gyrus (with a separate cluster in right anterior inferior frontal gyrus), bilateral presupplementary motor area (pre-SMA), bilateral retrosplenial cortex, left and right premotor cortex, left superior temporal sulcus, left intraparietal sulcus (IPS), right anterior insula, right posterior cingulate cortex (PCC), and right precuneus (see **Table 2** for MNI coordinates and statistics and **Figure 3** for locations and responses). Average betas for each of the stimulus type and repetition conditions were extracted from these clusters and subjected to follow-up analyses (repeated-measures ANOVAs and linear contrasts). The follow-up analysis revealed a significant linear trend of repetition (collapsing over stimulus type) in each of the clusters (*p*'s < 0.01), consistent with our hypotheses of sustained effects across numerous repetitions. All linear trends were negative except for the PCC and precuneus (see **Figure 3**, right panel). There was a main effect of stimulus type with greater activation for general software screenshots than security warnings (i.e., Business > Warning) in the left and right visual stream (dorsal and ventral), and the retrosplenial cortex (**Table 2**). The opposite effect (i.e., Warning > Business) was observed in clusters in the left inferior frontal gyrus, the pre-SMA, and the right anterior insula. The stimulus type by repetition number interaction was not significant in any cluster. Finally, there was a stimulus type by repetition number interaction in the linear trends in the left and right (dorsal) visual processing streams (**Table 2**).

The reduced activation with repeated exposures to stimuli may have represented participant failure to respond to stimuli or overall fatigue. As the behavioral orienting task was a subjective judgment, we were not able to calculate an accuracy rate to determine if accuracy decreased with the duration of the task. Nevertheless, response rates remained high (>94%) throughout the course of the task. As a check for overall fatigue, we modeled the single-presentation general computing screenshots. If the observed decreases in activation were due to overall fatigue, the effect should generalize to the novel stimuli as well. In all clusters of activation the activity for the novel stimuli was greater than for the final presentation of either the general computing or warning stimuli (**Figure 3**), with the sole exception of the warning stimuli in the right anterior insula.

The sustained negative linear trends in the majority of clusters of activation are consistent with habituation processes. Conversely, activation in the right precuneus increased with repeated exposures to the stimuli. The precuneus is a hub of the default mode network (DMN; Raichle, 2015), which is a network of brain structures that become more active as participants engage less in a primary task (Mason et al., 2007). To test whether the increasing activation observed in the precuneus represented DMN activation, we conducted a similarity analysis by extracting the mean betas for each condition in the precuneus cluster and calculating a correlation with this pattern of activation across every voxel in the brain. Correlation coefficients were Fisher transformed and a *t*-test was performed on these values versus 0 to identify regions where activation was significantly correlated with that of the precuneus. Five clusters were identified: the precuneus, posterior cingulate cortex, right temporal parietal junction, medial prefrontal cortex, and right frontopolar cortex (**Figure 4** and **Table 3**). As these regions are commonly associated with the DMN (Raichle, 2015), we conclude that the increasing activation with

TABLE 2 | Location and description of significant clusters showing a main effect of repetition.

			MNI coordinates			**ME stimulus type**			**ME repetition**			**INTX stim. type × repetition**			**Linear trend INTX**		
Label	**#Voxels**	**Direction**	***X***	***Y***	***Z***	***F*(1,21)**	***p***	η^2_p	***F*(5,105)**	***p***	η^2_p	***F*(5,105)**	***p***	η^2_p	***F*(1,21)**	***p***	η^2_p
L. Visual	483	Negative	−46	−59	−8	24.699	**<0.001**	0.54	26.391	**<0.001**	0.557	2.113	0.07	0.091	4.791	**0.04**	0.186
R. Dorsal visual	213	Negative	29	−66	34	10.953	**0.003**	0.343	26.499	**<0.001**	0.558	2.23	0.057	0.096	5.566	**0.028**	0.21
R. Ventral visual	173	Negative	29	−80	−11	7.469	**0.012**	0.262	24.083	**<0.001**	0.534	1.992	0.086	0.087	2.745	0.112	0.116
L. Inferior frontal gyrus	124	Negative	−40	10	27	30.564	**<0.001**	0.593	24.964	**<0.001**	0.543	0.394	0.852	0.018	0.337	0.567	0.016
B. Retrosplenial cortex	90	Negative	−5	−52	16	87.007	**<0.001**	0.806	13.861	**<0.001**	0.398	1.615	0.162	0.071	3.069	0.094	0.128
B. Pre-SMA	58	Negative	−9	16	64	21.41	**<0.001**	0.505	13.054	**<0.001**	0.383	1.496	0.198	0.066	1.274	0.272	0.057
R. Inferior frontal gyrus	32	Negative	40	6	27	3.03	0.096	0.126	15.056	**<0.001**	0.418	1.396	0.232	0.062	1.151	0.296	0.052
L. Premotor cortex	32	Negative	−29	16	54	2.935	0.101	0.123	21.341	**<0.001**	0.504	0.407	0.843	0.019	0.228	0.638	0.011
L. Superior temporal sulcus	21	Negative	−53	−4	−15	1.338	0.26	0.06	21.313	**<0.001**	0.504	0.453	0.81	0.021	0.128	0.724	0.006
R. Premotor cortex	20	Negative	33	−4	58	0.019	0.891	0.001	11.598	**<0.001**	0.367	1.264	0.286	0.059	2.622	0.121	0.116
R. Anterior inferior frontal gyrus	19	Negative	40	30	23	1.792	0.195	0.079	11.098	**<0.001**	0.346	0.29	0.917	0.014	0.154	0.698	0.007
L. Intraparietal sulcus	15	Negative	−29	−59	54	3.143	0.091	0.13	14.602	**<0.001**	0.41	0.141	0.982	0.007	0.505	0.485	0.023
R. Anterior insula	14	Negative	29	27	3	17.297	**<0.001**	0.464	15.996	**<0.001**	0.444	0.571	0.722	0.028	0.05	0.826	0.002
R. Posterior cingulate cortex	14	Positive	5	−25	30	0.997	0.329	0.045	10.875	**<0.001**	0.341	0.2	0.962	0.009	0.469	0.501	0.022
R. Precuneus	13	Positive	9	−70	37	1.076	0.311	0.049	0.9114	**<0.001**	0.303	0.65	0.662	0.03	1.589	0.221	0.07

Significant effects ($p < 0.05$) are indicated in boldface. L, Left; R, Right; B, Bilateral; ME, Main Effect; INTX, Interaction.

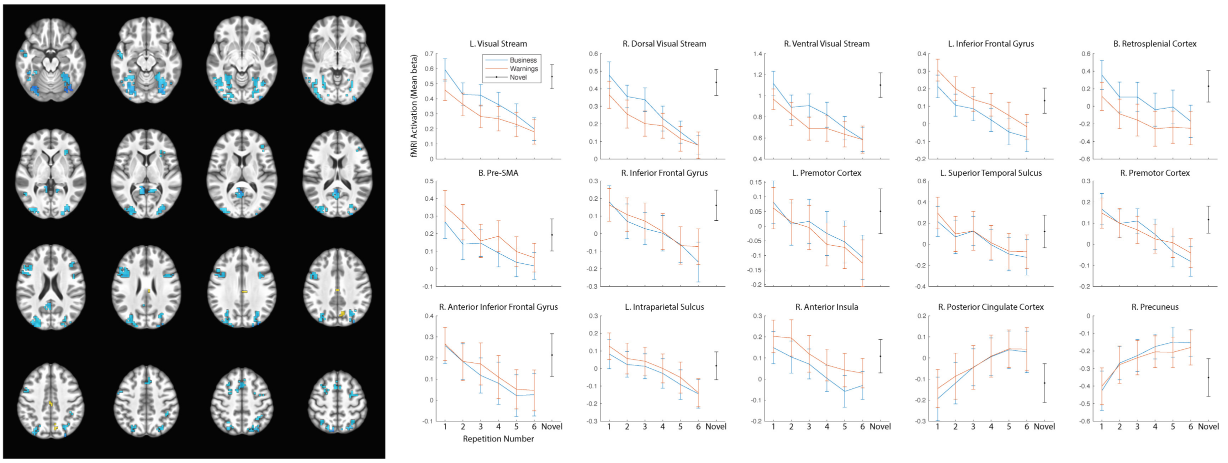


FIGURE 3 | Significant clusters of activation demonstrating a main effect of repetition number displayed on axial slices (left). Regions with negative linear trends are indicated in cooler colors (blue) while regions with positive linear trends are indicated in warmer colors (yellow/orange). Mean betas for the general software and security warning conditions over six repetitions within each cluster are displayed on the right. Posterior clusters in the visual processing streams, and retrosplenial cortex displayed a main effect of stimulus type with greater activation for the general software images than security warnings. More anterior regions including the left inferior frontal gyrus, the pre-SMA, and the right anterior insula displayed a main effect of stimulus type with greater activation for security warnings than general software images. The linear trend was significant in all clusters, however only the PCC and precuneus displayed a positive trend. Activation for the single-presentation general computing screenshots was significantly different from the final presentation of the repeated images in each cluster for both conditions with the exception of the Warning stimuli in the right anterior insula. Error bars, SEM.

TABLE 3 | Clusters significantly correlated with activation in the precuneus.

Label	Voxels	MNI coordinates		
		X	*Y*	*Z*
B. Precuneus	265	9	−70	37
R. Temporal parietal junction	118	50	−52	40
B. Posterior cingulate cortex	76	2	−28	27
R. Frontopolar cortex	46	22	58	−8
B. Medial prefrontal cortex	42	2	40	13

L, Left; R, Right; B, Bilateral.

stimulus repetition observed in the precuneus likely reflects increased DMN activation.

DISCUSSION

In the current experiment, participants viewed repeated images of software applications and security warnings while they underwent fMRI. We found evidence of repetition suppression for both stimulus types throughout the visual processing stream. Critically, fMRI activation continued to decrease over all six repetitions of the stimuli, indicating a continued repetition suppression effect with continued stimulus exposures. Conversely, we observed increased activation in DMN regions with repeated exposures. Finally, we observed increased activation in frontal regions including the pre-SMA, inferior frontal gyrus, and anterior insula for security warning stimuli compared to general software applications, consistent with heightened negative subjective value for the warning stimuli. These findings indicate that repetition suppression is multifaceted, differentially affecting a variety of areas.

We first examined the repetition suppression effect to everyday stimuli. We observed distinct patterns of activation over the course of repetitions. Similar to previous studies, there was a decrease of activation in areas related to visual processing, namely in the occipital lobe (Kovacs et al., 2013; Mayrhauser et al., 2014) and inferior temporal lobe (Summerfield et al., 2008). Adding to this previous work, we observed a continued, linear decrease in activation through all six repetitions. Such a finding shows that the decrease of activation in these areas does not level off after the second trial but continues to decrease with prolonged exposure to the stimulus. This repetition suppression occurred in frontal regions as well and applied to both images of general computing software and security warnings, indicating that the learned negative valence of computer security warnings is not enough to overcome habituation.

Along with the decreased activation in the occipital and inferior temporal lobes with repeated presentations, we also observed increased activation in the DMN, namely the precuneus and PCC. Activation in the DMN has been demonstrated to be negatively correlated with activation in a network of regions known to be involved in directing external attention, the dorsal attention network (Fox et al., 2005). Thus, increased activation in the DMN is often associated with unconstrained mental activity ("mind wandering") (Mason et al., 2007; Raichle and Snyder, 2007). The continued increased activation in the DMN during subsequent stimulus presentations suggests that the participants were less attentive to the stimuli as the repetitions increased.

Because we used naturalistic stimuli with learned negative valence, we were able to examine the differential response to positive and negative valance images over several repetitions. We observed greater activation for general software stimuli than the warning stimuli in posterior regions including the bilateral visual stream and retrosplenial cortex. The general software images were on average larger (general software mean image dimensions: 760.8 × 1,173.4 pixels; warning mean image dimensions: 381.9 × 589.4 pixels), which might have accounted for some of the greater activation in the visual processing stream. To control for this, we entered stimulus

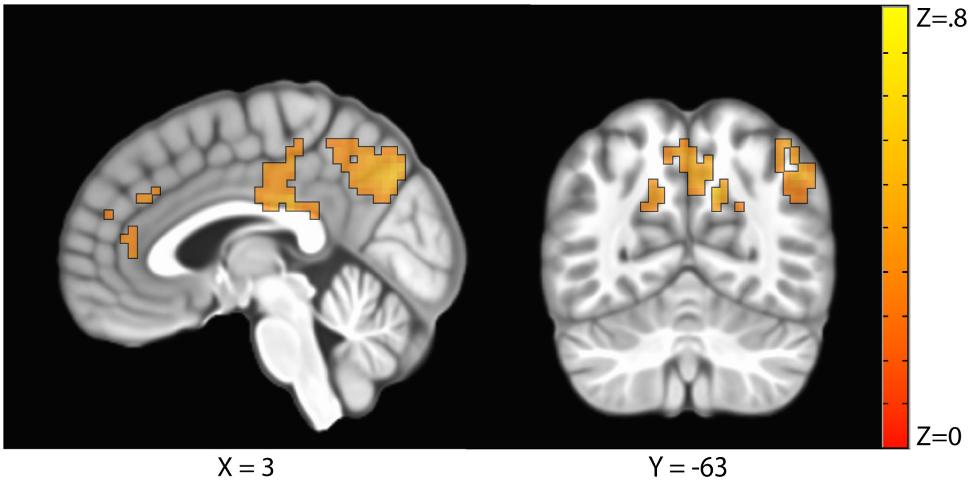


FIGURE 4 | Clusters where activation was significantly correlated with the precuneus in a representational similarity analysis (RSA) included the precuneus, the posterior cingulate cortex, the right temporal parietal junction, the medial prefrontal cortex, and the right frontopolar cortex.

size on each trial as a nuisance regressor in the first-level regression analysis. In spite of this control, we nevertheless observed widespread activation differences between stimulus types in the visual processing streams. This could be explained by elements of the images as the general software stimuli contained images used for work and recreation providing various uses, options, and tools. In contrast, the security warnings were less captivating with a lack of information and visual stimuli within the image. In spite of this, the linear trend interaction between stimulus types in the visual processing streams and intraparietal sulcus indicates that any additional visual or attentional processing afforded the general computing images habituated faster (i.e., had a steeper negative linear trend) than security warnings.

For areas including inferior frontal gyrus, pre-SMA, and right anterior insula, there was a greater level of activation for the computer warning stimuli than the general software images. The anterior insula has been specifically associated with anxiety and fear conditioning (Grupe and Nitschke, 2013) and has been implicated in initiating a fear response as a result from negative or harmful stimuli (Knutson et al., 2014). The greater activation for computer warning stimuli as opposed to general software images in this region is consistent with a fear response to the warning stimuli over the general software stimuli. An anterior insula-mediated fear response functions not only for environmental risk, but also for safety from other negative experiences and stimuli (Knutson et al., 2014). Additionally, the anterior insula response was still activated even though the computer warning stimuli was fictitious. The participant was informed before participation that the computer warnings were mock images and not directly related to them or their property. This is consistent with other studies that have shown that fear response is still activated even when not part of the primary task (Carlsson et al., 2004).

Some limitations should be noted in the present study. First, while we examined the repetition suppression effect with complex stimuli, we looked at these stimuli with repeated repetitions within a short period of time. The use of complex stimuli adds to the external validity of the study, but computer security warnings are generally observed infrequently over longer periods of time (days or weeks). A longitudinal study looking at how extended exposure over several weeks could add to the findings of this study by presenting these stimuli in a more natural time course. Second, computer security warnings are a familiar sight among individuals who regularly use computers. Further, we did not assess the pre-experimental familiarity of the stimuli in this group of participants. Therefore, these stimuli may not have been completely novel. Regardless, we still found a strong repetition suppression effect even when the participants had encountered similar stimuli previously in everyday use of computers. This suggests a potential line of research examining the extent to which habituation generalizes from non-security messages to computer security warnings. In other words, future studies may wish to examine whether participants habituate to innocuous system notifications (such as email notifications) and whether that habituation generalizes to security warnings. Third, we do not determine the number of repetitions where activation begins to level off. While other research shows that the greatest decrease in activation occurs during early repeated exposures to stimuli to complex (Anderson et al., 2016a,b; Vance et al., 2018), future research is needed to determine at what point additional repetitions do not cause a meaningful decrease in activation. Finally, we did not collect valance ratings or physiological arousal measurements associated with the warning stimuli. However, previous studies (e.g., Buck et al., 2017) have demonstrated negative valance associated with pop-up security warnings.

One strength of this study is that it examined the repetition suppression effect in complex, everyday stimuli as well as examining this phenomenon with extended repetitions. This design allowed us to replicate and confirm previous findings of earlier research that visual processing activation decreases over repetition as well as DMN activation increases over repetition. Along with confirming prior research on the subject, the use of complex stimuli allows these findings to be generalized to greater variations of stimuli than have been used in prior research. Finally, we demonstrated that the anterior insula responded to the negative valence of the computer warning stimuli and that this increased activation also demonstrated a repetition suppression effect over continued exposures. The habituation to warnings is a concern for computer security as users are less likely to attend and respond appropriately to repeated computer security warnings.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available at https://doi.org/10.18112/openneuro.ds002363.v1.1.1. The code used in the analyses presented is available at https://github.com/Kirwanlab/RepetitionSuppression.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Brigham Young University Institutional Review Board. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

CBK, BA, AV, DE, and JJ conceptualized the experiment. DB, BA, AV, DE, and JJ collected stimuli and designed the experimental paradigm. CBK, DB, BA, and AV collected the data. CBK and DB performed the data analysis. CBK, DB, BA, AV, DE, and JJ edited and wrote the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the National Science Foundation (1422831 to BA and 1247842 to DE) and Google (FRA-194 to AV).

REFERENCES

Akhawe, D., and Felt, A. P. (2013). "Alice in warningland: a large-scale field study of browser security warning effectiveness," in *Proceedings of the 22nd USENIX Conference on Security*, Washington, DC.

Anderson, B. B., Jenkins, J. L., Vance, A., Kirwan, C. B., and Eargle, D. (2016a). Your memory is working against you: how eye tracking and memory explain habituation to security warnings. *Dec. Support Syst.* 92, 3–13.

Anderson, B. B., Vance, A., Kirwan, C. B., Jenkins, J., and Eargle, D. (2016b). From warnings to wallpaper: why the brain habituates to security warnings and what can be done about it. *J. Manag. Inform. Syst.* 33, 713–743.

Berlyne, D. E. (1970). Novelty, complexity, and hedonic value. *Percept. Psychophys.* 8, 279–286.

Bravo-Lillo, C., Cranor, L., Komanduri, S., Schechter, S., and Sleeper, M. (2014). "Harder to ignore? Revisiting pop-up fatigue and approaches to prevent it," in *Proceedings of the 10th Symposium On Usable Privacy and Security ({SOUPS} 2014)*, Menlo Park, CA.

Buchel, C., and Dolan, R. J. (2000). Classical fear conditioning in functional neuroimaging. *Curr. Opin. Neurobiol.* 10, 219–223.

Buck, R., Khan, M., Fagan, M., and Coman, E. (2017). The user affective experience scale: a measure of emotions anticipated in response to pop-up computer warnings. *Int. J. Hum. Comput. Interact.* 34, 25–34. doi: 10.1080/10447318.2017.1314612

Carlsson, K., Petersson, K. M., Lundqvist, D., Karlsson, A., Ingvar, M., and Ohman, A. (2004). Fear and the amygdala: manipulation of awareness generates differential cerebral responses to phobic and fear-relevant (but nonfeared) stimuli. *Emotion* 4, 340–353. doi: 10.1037/1528-3542.4.4.340

Chouinard, P. A., Morrissey, B. F., Kohler, S., and Goodale, M. A. (2008). Repetition suppression in occipital-temporal visual areas is modulated by physical rather than semantic features of objects. *Neuroimage* 41, 130–144. doi: 10.1016/j.neuroimage.2008.02.011

Costa-Faidella, J., Baldeweg, T., Grimm, S., and Escera, C. (2011). Interactions between "what" and "when" in the auditory system: temporal predictability enhances repetition suppression. *J. Neurosci.* 31, 18590–18597. doi: 10.1523/jneurosci.2599-11.2011

Cox, R. W. (1996). AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Comput. Biomed. Res.* 29, 162–173. doi: 10.1006/cbmr.1996.0014

Cox, R. W., Chen, G., Glen, D. R., Reynolds, R. C., and Taylor, P. A. (2017). FMRI clustering in AFNI: false-positive rates redux. *Brain Connect.* 7, 152–171.

Danckert, S. L., Gati, J. S., Menon, R. S., and Kohler, S. (2007). Perirhinal and hippocampal contributions to visual recognition memory can be distinguished from those of occipito-temporal structures based on conscious awareness of prior occurrence. *Hippocampus* 17, 1081–1092. doi: 10.1002/hipo.20347

Daselaar, S. M., Fleck, M. S., and Cabeza, R. (2006). Triple dissociation in the medial temporal lobes: recollection, familiarity, and novelty. *J. Neurophysiol.* 96, 1902–1911. doi: 10.1152/jn.01029.2005

Desmond, J. E., and Glover, G. H. (2002). Estimating sample size in functional MRI (fMRI) neuroimaging studies: statistical power analyses. *J. Neurosci. Methods* 118, 115–128.

Dimoka, A. (2012). How to conduct a functional magnetic resonance (fMRI) study in social science research. *MIS Q.* 36, 811–840.

Fox, M. D., Snyder, A. Z., Vincent, J. L., Corbetta, M., Van Essen, D. C., and Raichle, M. E. (2005). The human brain is intrinsically organized into dynamic, anticorrelated functional networks. *Proc. Natl. Acad. Sci. U.S.A.* 102, 9673–9678. doi: 10.1073/pnas.0504136102

Grill-Spector, K., Henson, R., and Martin, A. (2006). Repetition and the brain: neural models of stimulus-specific effects. *Trends Cogn. Sci.* 10, 14–23. doi: 10.1016/j.tics.2005.11.006

Grotheer, M., and Gyula, K. (2015). The relationship between stimulus repetitions and fulfilled expectations. *Neuropsychologia* 67, 175–182. doi: 10.1016/j.neuropsychologia.2014.12.017

Grupe, D. W., and Nitschke, J. B. (2013). Uncertainty and anticipation in anxiety: an integrated neurobiological and psychological perspective. *Nat. Rev. Neurosci.* 14, 488–501. doi: 10.1038/nrn3524

Kirwan, C. B., Shrager, Y., and Squire, L. R. (2009). Medial temporal lobe activity can distinguish between old and new stimuli independently of overt behavioral choice. *Proc. Natl. Acad. Sci. U.S.A.* 106, 14617–14621. doi: 10.1073/pnas.0907624106

Knutson, B., Katovich, K., and Suri, G. (2014). Inferring affect from fMRI data. *Trends Cogn. Sci.* 18, 422–428. doi: 10.1016/j.tics.2014.04.006

Kovacs, G., Kaiser, D., Kaliukhovich, D. A., Vidnyanszky, Z., and Vogels, R. (2013). Repetition probability does not affect fMRI repetition suppression for objects. *J. Neurosci.* 33, 9805–9812. doi: 10.1523/jneurosci.3423-12.2013

Lacy, J. W., Yassa, M. A., Stark, S. M., Muftuler, L. T., and Stark, C. E. (2011). Distinct pattern separation related transfer functions in human CA3/Dentate and CA1 revealed using high-resolution fmri and variable mnemonic similarity. *Learn. Mem.* 18, 15–18.

Larsson, J., and Smith, A. T. (2012). fMRI repetition suppression: neuronal adaptation or stimulus expectation? *Cereb. Cortex* 22, 567–576. doi: 10.1093/cercor/bhr119

Li, X., Morgan, P. S., Ashburner, J., Smith, J., and Rorden, C. (2016). The first step for neuroimaging data analysis: DICOM to NIfTI conversion. *J. Neurosci. Methods* 264, 47–56.

Malmierca, M. S., Cristaudo, S., Perez-Gonzalez, D., and Covey, E. (2009). Stimulus-specific adaptation in the inferior colliculus of the anesthetized rat. *J. Neurosci.* 29, 5483–5493. doi: 10.1523/jneurosci.4153-08.2009

Mason, M. F., Norton, M. I., Van Horn, J. D., Wegner, D. M., Grafton, S. T., and Macrae, C. N. (2007). Wandering minds: the default network and stimulus-independent thought. *Science* 315, 393–395. doi: 10.1126/science.1131295

Mayrhauser, L., Bergmann, J., Crone, J., and Kronbichler, M. (2014). Neural repetition suppression: evidence for perceptual expectation in object-selective regions. *Front. Hum. Neurosci.* 8:225. doi: 10.3389/fnhum.2014.00225

McDonald, C. R., Thesen, T., Carlson, C., Blumberg, M., Girard, H. M., and Trongnetrpunya, A. (2010). Multimodal imaging of repetition priming: using fMRI, MEG, and intracranial EEG to reveal spatiotemporal profiles of word processing. *Neuroimage* 53, 707–717. doi: 10.1016/j.neuroimage.2010.06.069

Muncy, N. M., Hedges-Muncy, A. M., and Kirwan, C. B. (2017). Discrete pre-processing step effects in registration-based pipelines, a preliminary volumetric study on T1-weighted image. *PLoS One* 12:e0186071. doi: 10.1371/journal.pone.0186071

Olsson, A., and Phelps, E. A. (2007). Social learning of fear. *Nat. Neurosci.* 10, 1095–1102. doi: 10.1038/nn1968

Paulus, M. P., and Stein, M. B. (2006). An insular view of anxiety. *Biol. Psychiatry* 60, 383–387. doi: 10.1016/j.biopsych.2006.03.042

Pechmann, C., and Stewart, D. W. (1988). Advertising repetition: a critical review of wearin and wearout. *Curr. Issues Res. Advert.* 11, 285–329. doi: 10.1080/01633392.1988.10504936

Phelps, E. A., O'Connor, K. J., Gatenby, J. C., Gore, J. C., Grillon, C., and Davis, M. (2001). Activation of the left amygdala to a cognitive representation of fear. *Nat. Neurosci.* 4, 437–441. doi: 10.1038/86110

Phillips, R. G., and LeDoux, J. E. (1992). Differential contribution of amygdala and hippocampus to cued and contextual fear conditioning. *Behav. Neurosci.* 106, 274–285.

Raichle, M. E. (2015). The brain's default mode network. *Annu. Rev. Neurosci.* 38, 433–447.

Raichle, M. E., and Snyder, A. Z. (2007). A default mode of brain function: a brief history of an evolving idea. *Neuroimage* 37, 1083–1090. doi: 10.1016/j.neuroimage.2007.02.041

Schmidt, S., and Eisend, M. (2015). Advertising repetition: a meta-analysis on effective frequency in advertising. *J. Advert.* 44, 415–428.

Segaert, K., Weber, K., de Lange, F. P., Petersson, K. M., and Hagoort, P. (2013). The suppression of repetition enhancement: a review of fMRI studies. *Neuropsychologia* 51, 59–66.

Shrager, Y., Kirwan, C. B., and Squire, L. R. (2008). Activity in both hippocampus and perirhinal cortex predicts the memory strength of subsequently remembered information. *Neuron* 59, 547–553. doi: 10.1016/j.neuron.2008.07.022

Stern, C. E., Corkin, S., Gonzalez, R. G., Guimaraes, A. R., Baker, J. R., and Jennings, P. J. (1996). The hippocampal formation participates in novel picture encoding: evidence from functional magnetic resonance imaging. *Proc. Natl. Acad. Sci. U.S.A.* 93, 8660–8665. doi: 10.1073/pnas.93.16.8660

Summerfield, C., Trittschuh, E. H., Monti, J. M., Mesulam, M. M., and Egner, T. (2008). Neural repetition suppression reflects fulfilled perceptual expectations. *Nat. Neurosci.* 11, 1004–1006. doi: 10.1038/nn.2163

Summerfield, C., Wyart, V., Johnen, V. M., and de Gardelle, V. (2011). Human scalp electroencephalography reveals that repetition suppression varies with expectation. *Front. Hum. Neurosci.* 5:67. doi: 10.3389/fnhum.2011.00067

Todorovic, A., and de Lange, F. P. (2012). Repetition suppression and expectation suppression are dissociable in time in early auditory evoked fields. *J. Neurosci.* 32, 13389–13395. doi: 10.1523/jneurosci.2227-12.2012

Todorovic, A., van Ede, F., Maris, E., and de Lange, F. P. (2011). Prior expectation mediates neural adaptation to repeated sounds in the auditory cortex: an MEG study. *J. Neurosci.* 31, 9118–9123. doi: 10.1523/jneurosci.1425-11.2011

Valentini, E. (2011). The role of perceptual expectation on repetition suppression: a quest to dissect the differential contribution of probability of occurrence and event predictability. *Front. Hum. Neurosci.* 5:143. doi: 10.3389/fnhum.2011.00143

Vance, A., Jenkins, J. L., Anderson, B. B., Bjornn, D., and Kirwan, B. (2018). Tuning out security warnings: a longitudinal examination of habituation through fMRI, eye tracking, and field experiments. *MIS Q.* 42, 355–380.

Vance, A., Kirwan, B., Bjornn, D., Jenkins, J., and Anderson, B. B. (2017). "What do we really know about how habituation to warnings occurs over time? A longitudinal fMRI study of habituation and polymorphic warnings," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, (New York, NY: Association for Computing Machinery).

Wogalter, M., and Vigilante, W. J. Jr. (2006). "Attention switch and maintenance," in *Handbook of Warnings*, ed. M. S. Wogalter (Mahwah, NJ: Erlbaum), 245–266.

Conflict of Interest: The authors declare that this study received funding from Google. The funder was not involved in the study design, collection, analysis, interpretation of data, the writing of this article or the decision to submit it for publication.

HYPOTHESIS AND THEORY
published: 16 December 2020
doi: 10.3389/fpsyg.2020.537612

Interdisciplinary Lessons Learned While Researching Fake News

Char Sample[1]*, Michael J. Jensen[2], Keith Scott[3], John McAlaney[4], Steve Fitchpatrick[5], Amanda Brockinton[4]*, David Ormrod[5] and Amy Ormrod[5]

[1] Idaho National Laboratory, Idaho Falls, ID, United States, [2] Institute for Governance and Policy Analysis, University of Canberra, Canberra, ACT, Australia, [3] Department of Linguistics, De Montfort University, Leicester, United Kingdom, [4] Department of Psychology, Bournemouth University, Bournemouth, United Kingdom, [5] Terra Schwartz, Canberra, ACT, Australia

OPEN ACCESS

Edited by:
Paul Watters,
La Trobe University, Australia

Reviewed by:
Kai Shu,
Illinois Institute of Technology,
United States
Chris Baber,
University of Birmingham,
United Kingdom

***Correspondence:**
Char Sample
charsample50@gmail.com
Amanda Brockinton
Abrockinton@bournemouth.ac.uk

Specialty section:
This article was submitted to
Cognition,
a section of the journal
Frontiers in Psychology

Received: 06 March 2020
Accepted: 06 October 2020
Published: 16 December 2020

Citation:
Sample C, Jensen MJ, Scott K,
McAlaney J, Fitchpatrick S,
Brockinton A, Ormrod D and
Ormrod A (2020) Interdisciplinary
Lessons Learned While Researching
Fake News.
Front. Psychol. 11:537612.
doi: 10.3389/fpsyg.2020.537612

The misleading and propagandistic tendencies in American news reporting have been a part of public discussion from its earliest days as a republic (Innis, 2007; Sheppard, 2007). "Fake news" is hardly new (McKernon, 1925), and the term has been applied to a variety of distinct phenomenon ranging from satire to news, which one may find disagreeable (Jankowski, 2018; Tandoc et al., 2018). However, this problem has become increasingly acute in recent years with the Macquarie Dictionary declaring "fake news" the word of the year in 2016 (Lavoipierre, 2017). The international recognition of fake news as a problem (Pomerantsev and Weiss, 2014; Applebaum and Lucas, 2016) has led to a number of initiatives to mitigate perceived causes, with varying levels of success (Flanagin and Metzger, 2014; Horne and Adali, 2017; Sample et al., 2018). The inability to create a holistic solution continues to stymie researchers and vested parties. A significant contributor to the problem is the interdisciplinary nature of digital deception. While technology enables the rapid and wide dissemination of digitally deceptive data, the design and consumption of data rely on a mixture of psychology, sociology, political science, economics, linguistics, marketing, and fine arts. The authors for this effort discuss deception's history, both old and new, from an interdisciplinary viewpoint and then proceed to discuss how various disciplines contribute to aiding in the detection and countering of fake news narratives. A discussion of various fake news types (printed, staged events, altered photographs, and deep fakes) ensues with the various technologies being used to identify these; the shortcomings of those technologies and finally the insights offered by the other disciplines can be incorporated to improve outcomes. A three-point evaluation model that focuses on contextual data evaluation, pattern spread, and archival analysis of both the author and publication archives is introduced. While the model put forth cannot determine fact from fiction, the ability to measure distance from fact across various domains provides a starting point for evaluating the veracity of a new story.

Keywords: fake news, discipline, behaviors, values, rhetoric, politics, deception

"If it is not true, it is very well invented." —Giordano Bruno

INTRODUCTION

Fake news has a long history in America (McKernon, 1925; Innis, 2007), becoming internationally recognized as a problem in 2016, the year it was declared word of the year by Macquarie Dictionary (Lavoipierre, 2017). The re-emergence of the term "fake news" (Meza, 2017) served as an inflection point for academics across various disciplines. Some academics observed the commonalities between "fake news" and propaganda that uses a different delivery mechanism (Younger, 2018), whereas others observed greater sophistication, customization, and weaponization (Younger, 2018; Verrall and Mason, 2019). Those in academia and government who recognized this threat, some as early as Szfranski (1997), others soon after (Cybenko et al., 2002). Szfranski (1997) suggested that the weaponization of deceptive information would require protection of both combatants and non-combatants alike. While some groups of people are more resilient against deceptive data (Bjola and Papadakis, 2020) suggesting a cultural component, significant populations remain vulnerable. The vulnerable also includes journalists, who repeat the stories that align with their own values. Even the journalists' verification and validation mechanisms are corrupted by algorithms that provide information that aligns with targeted beliefs. Indeed, according to some behavioral scientists, all are vulnerable to the messages that confirm biases (Oswald and Grosjean, 2004).

Media spheres such as journalism require their journalists to act as watchdogs of information-sharing for its global citizens. Their position in the world holds responsibility to provide independent truth and legitimacy to its audience by providing fact-checking. General verification procedures will be covered to transition into discussion about the importance of interdisciplinary work. Journalists intercept deception by reporting on the truth of our reality, having a general agreed normative approach to fact-checking including combating fake news narratives even if interpreting false claims is still very much subjective (Graves, 2018; Mena, 2019). Five elements of fact checking provided by Bratich (2020) include choosing claims to check, contacting the source of the claim, tracing false claims, working with experts, and showing their (journalists) work. Within this frame, journalism has become loaded with uncertainty, mistrust, and manipulation from its user engagement and many other trends, such as politics and emerging technologies, which intersect it (Waisbord, 2018).

Pomerantsev and Weiss (2014) identified five goals of disinformation, of which fake news is a subset, these goals include paralysis, demoralization, confusion, blackmail, and subversion. Disinformation campaigns will seek any and all of these goals each of the five goals represents a strategy to use against a targeted group of people. For example, credible news stories that report opposite stories on the same event can, if both are professionally done, confuse a person who is new to the story and environment, rather than simply sway the person to one side or another.

Fake News, Disinformation, and Manipulation

The use of disinformation and misinformation in news has a long history. Some scholars have focused on misinformation, the inadvertent release of misleading or factually incorrect information, or disinformation, involving the intentional diffusion of factually incorrect claims for political purposes (Bakir and McStay, 2018; Bennett and Livingston, 2018). From a perspective of information warfare, however, claims need not be false to have strategic value in manipulating an audience (Schafer, 2018). Well-timed, factually correct information can be as effective as a lie; when this occurs, the information becomes weaponized. This section develops the distinction between information and its weaponization as the movement from an information logic to an identity logic within communications.

Information is classically defined as "current data about developments in and the status of" a system at any given time (Downs, 1957, p. 79). An information logic has three components. First, there is a temporal dimension. As information refers to the current status of a system or scene of events, it has a fleeting duration, "information that is repeated is no longer information. It retains its meaning in the repetition but loses its value as information" because it no longer updates one's understanding of a state of affairs (Luhmann, 1995, p. 67; Lash, 2002). The same materials may resurface repeatedly, but this defines an extended present that sublimates past and future and is no longer operating in an informational mode (Rushkoff, 2013).

Second, the contextualized component includes the scene or situation in which activity takes place. Contextualized data are bound to the environment where created or resides. In this sense, data, particularly digital data, when taken out of context, perturb the environment, and traces from the perturbation will remain.

Information has a third characteristic: information descriptive, rather than moralizing or hortatory, does not call for action (Burke, 1969, p. 41). Thus, information contrasts with emotionally polarizing communication content. Whereas the experience of sensory data may provide an updated status about local states of affairs, emotions "do not give us any information about the world" (Wittgenstein, 1967, p. 491).

Both misinformation and disinformation point to informational disorders, an enduring aspect of the Downsean tradition within political science. Accurate information about states of affairs within a political system has remained a critical currency, often in too short supply, which citizens require to make informed democratic choices (Converse, 1962; Grofman, 1995; Delli et al., 1997; McGann, 2006). Fake news is often taken as a species of disinformation as it is both a fabrication and "mimics news media content in form but not in process or intent" (Lazer, 2018, p. 1094). From this perspective, misinformation and disinformation are distorting as these phenomena provide errant premises on which to make decisions in light of pregiven preferences (Hochschild and Katherine, 2015; Mathiesen, 2018). And, while deception might be a short-term strategy for elected officials, if they seek reelection, there can be costs to deception during elections (Ferejohn, 1990, p. 10). More recently, the rise of polarized electorates makes possible the continuation of informational disorders.

Preparing for Fake News

Deception is not limited to the political realm. Health/medicine (Springer, 2020), finance (Cybenko et al., 2002), the military, and cyber domains are a few of the other environments where deceptive data or fake news have successfully been deployed. Deception relies on tricking both sensation and perception. In order for information to be perceived the data must first be sensed. Human perception and information processing are still not fully understood (De Faveri et al., 2017).

"Persuasion lies at the heart of political communications" (Flanagin and Metzger, 2014, p. 1). In order for fake news to be effective, the deception must first be formed then communicated. The communication process where sender and receiver share the same perception of information was defined by Gerbner (1956). Gerbner's model of communications identifies the role of perception and contextualization in the message creation phase. Successful deception relies on influencing the thinking of the target (De Faveri et al., 2017). The target in the case of fake news is the human mind, and journalists are humans and thus vulnerable to disinformation.

Because the target is the human mind, a brief discussion on factors that feed into decision-making is relevant. Before information processing can occur, the data that form the information must first be sensed and then perceived. Sensing occurs when one or more of the five senses are stimulated. Generally speaking, sight and sound predominate in sensory stimulation particularly, so in the online world where fake news prevails, sensory deprivation makes possible the ability to control perception. Perception provides the input into decision-making, so a lack of stimuli (sensation data) or manipulated sensation data designed as context triggers the initial unconscious, neural response for specific actions (Gazzaniga, 2014). Conversely, an un-sensed event is never perceived or a non-event.

While much focus on decision-making centers on motivation, the authors would be remiss if they did not list additional factors. Decisions rely on several factors that in addition to motivation include patience/impatience, risk attitude, and ambiguity attitude (Gazzaniga, 2014). These different factors suggest that the manner in which decisions are made is diverse. Any single factor or combination of factors and predicting which factors dominate in any decision are most accurate as a *post hoc* exercise. Furthermore, all factors discussed in the decision-making process are situationally dependent.

Decisions can be broken down into two types: conscious, otherwise known as action–outcome decisions, and unconscious, also referred to as stimulus–response (Dijksterhuis, 2004; Kahneman, 2011; Gazzaniga, 2014). Both types of decisions are influenced by biases. De Faveri et al. (2017) described patterned deviations from fact in perception as biases and heuristics—cognitive shortcuts (McAlaney and Benson, 2020). Regardless of the type of decision being made, the biases still influence. However, the message creation can vary.

De Faveri et al. (2017) examined deception in the hostile cyber environment and described three groups of biases: personal and cultural, organizational, and cognitive. It was noted that while highly effective, biases are difficult to obtain (ibid). However, this observation preceded widespread knowledge of the role of social media in fake news targeting and dissemination (Shu et al., 2017 several sources).

The new role of social media in content creation, delivery, and dissemination of fake news has changed the landscape in unanticipated ways, requiring a reexamination of ways to identify and ultimately mitigate this type of deception. For this reason, the authors are examining some of the common disciplines involved in fake news or digital deception within the model describing content, distribution, and archives defined by Sample et al. (2018).

Deception as a Strategy

The assumption and assertion of a paper such as this are the role of fake news as a means to achieve political, social, and potentially other forms of influence utilizing deception as a strategy. Deception as strategy has roots in ancient human behavior, observed in the earliest histories including Greek mythology (Phaedrus, 2008, p. 438). In more recent times, Erfurth's treatise on Surprise (Erfurth, 1943) provides a number of helpful insights. He observes that almost all decisive military victories have been preceded by surprise, which relies on secrecy and speed. Deception is a form of surprise, providing a means to unbalance an opponent through uncertainty. Handel's detailed analysis of deception at the strategic and operational levels in World War II also offers key observations. Deception must be believable to the target audience, with sufficient resources and time invested in a coherent narrative to reinforce existing beliefs: "The susceptibility to conditioning is one of the most fundamental human proclivities to be exploited by deception operations" (Handel, 1987, p. 14). Conditioning greatly precedes the actual event of deception. Conditioning lays the groundwork upon which the deception capitalizes.

Disinformation, a critical form of and enabler of deception, has a history in both warfare and state security functions. The use of disinformation as a form of deception is examined by Whaley (2007, p. 8), in the historical context that it was originally a World War I term applied by the German General Staff and then adopted by the Russians. Applying Shannon's communication model, relevant but false information is fed into a communication channel, forming a third transmission category to signal and noise. This third category described by Whaley (ibid) may be disinformation or misinformation depending on intent. Misinformation is inadvertent, whereas disinformation deliberately seeks to overload, discredit, or realign an audience's information management capabilities. Given the requirement to consider intent, disinformation has little utility without a purpose. Having understood the target and obtaining a means to access information and information networks and then subsequently exploiting those networks to expand access, disinformation provides the means to utilize supporting conditioned biases and narratives with the intent to influence perceptions and behaviors (Waltz, 2008, p. 4).

The Soviet concept of *maskirovka* probably best encapsulates the complexity of the problem space surrounding deception. Although the term *maskirovka* can be defined as camouflage, it extends in Soviet doctrine across a broad array of strategic, operational, and tactical measures to obscure intent, maintain security, and confuse the adversary (Glantz, 2006, p. 2). While there are numerous instances of these maskirovka strategies being successfully employed throughout World War II, which have helped inform modern doctrines and techniques, it is also important to note that failures have occurred when maskirovka was employed hastily, poorly coordinated, enacted by personnel with inadequate training (ibid, p. 14), or conducted in a stereotyped or patterned manner (ibid, p. 10).

A critical observation of the maskirovka concept is the employment of a variety of techniques at all levels, in a planned and coordinated manner that also sought to embrace complexity with significant focus on aligning tactical outcomes with strategic intent. This appears to be a commonality with today's use of fake news. The employment of various tactical and operational approaches to achieve a broad strategic intent allows for multiple target audiences to be engaged with sometimes conflicting narratives and thematic episodes. This tactical and operational flexibility could be regarded as dangerous and counterintuitive from the strategic perspective, but it provides freedom of maneuver across the information environment and the ability to leverage the complexity of the modern information environment to achieve specific outcomes efficiently and with speed. The ability to roam widely, engaging with numerous audiences, themes, and narratives, at speed, appears to be a force multiplier in the employment of disinformation and fake news through social media and online forums. Moreover, this approach capitalizes on a number of perceived failures by trusted agencies to apply their moral and ethical narratives consistently, meaning that conflicting narratives by fake news agencies can always be excused by way of pointing to inconsistencies by previously trusted establishments who have perceived conflicts of interest, often amplified by the same fake news outlets.

The identification of target variables and ability to fashion-specific messages at an individual level, refined based on their personal and cultural data, appear to support the contention that a single, cohesive narrative is not required in the modern world. Fake news agents, marketers, and political organizations are able to target specific individuals based on individual data collected from internet fingerprinting and social media. This target variable data can distinguish at an individual level likely biases, beliefs, and likely actions through personality profiles. The more traditional media, government, and military refer to target audiences in a different, less precise manner, based on broad narratives and a focus on broad beliefs and groupings with an assumption that these descriptions will lead to specific group behaviors. It appears from these differences and the rapid evolution of these technologies that disinformation campaigns have a distinct advantage in the modern information environment. In the instance where one can focus on the issues specific to each individual and fashion the message to alter behavior around those issues at a personal and granular level, it appears that the narrative can be delivered in a micro, targeted manner. The alternative appears to be the delivery of grand narratives and themes supported by "trusted" agencies that rely on their self-perception of impartiality, which is quickly a target for fake news agencies and those who are likely to benefit from distrust of alternatives to the fake news narratives. It remains to be seen if the employment of fact checking, controlled narratives, and traditional information operations approaches is sufficient for the information environment of the future, but the results to date are not particularly positive. Perhaps part of the problem is the inconsistency inherent in modern life—it is not inconceivable to act contrary to one's beliefs based on more personal, pertinent matters, which are fleeting. That is a matter priests, theologians, and ethicists have grappled with since the dawn of organized religion. Personality and culture, discussed later within this article, are factors that are likely to contribute to these outcomes.

The modern context of disinformation as applicable to fake news extends from the fundamental concepts of deception as a strategy and some of the principles discussed above. Susceptibility to conditioning, bias, narrative, and the exploitation of information and social networks are all fundamental to the concept of fake news. These concepts will be discussed in more detail throughout this article.

BACKGROUND

The weaponization of information is enhanced in the digital world where communities are created not only through borders and national boundaries, but also by shared thoughts that include shared hopes and fears (Bennett, 2012). In the information age where decision-making, especially in Western-style democracies, carries great importance, the ability to control sensing and manipulate perception in online communities is extremely valuable.

If war is political, and politics inhibits a variety of attributes of war, modern politics is in many ways invested in the preparation for war, and our existing politics may even stem from and reproduce a set of relationships established through war (Virilio and Sylvère, 2008). The existence of nuclear deterrence as a variation of the "stability-instability paradox" may incentivize subkinetic forms of warfare, which can be harder to deter precisely because efforts to restore deterrence conventionally can risk nuclear escalation (Gartzke and Lindsay, 2019, p. 14). Clausewitz (1982) termed war as politics by other means. However, it is equally possible in the modern context of hybrid war, political warfare, gray-zone conflict, and the like; politics may be a continuation of warfare by other means (Foucault, 2003).

Information warfare weaponizes communications in order to effect change in a target audience in terms of their attitudes and behaviors. If content is the currency of propaganda, then timing performs a similar function for information weaponization. The strategic communication of information can be a "source multiplier" shaping one's understanding of situations, as well as shaping "the operational environment" so as to neutralize an adversary, as well as advance one's own strategic objectives (Armistead, 2004, p. 1). While information warfare retains the descriptor "information," it denotes a field of communication that is transformative more than informative. Information operations achieve these ends by seeking to "influence, disrupt, corrupt, or usurp the decision making of adversaries" while protecting those capacities for one's own side. Because decision making is largely biased, and biases are behavioral in nature, the shaping of attitudes and beliefs is key to success in this environment.

Although information warfare has often been conceptually confined to a space of military warfare, there is growing recognition that it "can take place in any situation across the spectrum of war or peace" (Morgan and Thompson, 2018, p. 10) whereby warfare extends into political life in a non-kinetic, non–physically violent form (Singer and Brooking, 2018, sec. Kindle: 325). This new type of warfare has been referred to as hybrid warfare (Commin and Filiol, 2013) and enacted in numerous countries (Atkinson, 2018), where fake news as a weapon of information warfare plays a prominent role (Younger, 2018).

Politics have become a problematic center in fake news reporting; journalists are criticized for being non-partisan during fact-checking procedures coupled by general misunderstandings that they are responsible for fact-checking future statements from politicians, leading to increased user distrust in mainstream media (Uscinski, 2015). These trends have caused a shift in the traditional hierarchical information sphere on how truth is reported, and interdisciplinary work has the opportunity to address some of these issues. Social media has transformed the landscape of information reporting, meaning that solutions to combat fake news depend on the flexibility of traditional journalistic pathways to produce fact-checking frameworks clear enough to account for such change. Much like cybersecurity issues faced today, the information flow of fake news is unprecedented and at times overwhelming; mounting pressures on journalists to discern truths has allowed for both increased verification and vulnerabilities to occur in reporting. For example, fact-checking the credibility of sources is a common task, but there are many examples in mainstream media where this is not the case, and many journalists recognize the presence of this information disorder (Plotkina et al., 2020). Information excess from virtual space produces a reporting experience that can cause fact-checking to be synonymous with instantaneous discernment of sources to then act, or not act, upon. Generally, journalists do agree that taking the time to do fact-checking thoroughly is more important than being the first to cover a story, but this is not always a tangible result (Schapals, 2018). Journalism, unlike many other spheres, have an advantage in its ability to report on evolving fake news concepts at the rate fake news articles are being produced because of its unique access to information streams.

There is an alternative view that modern information operations occur in a globally competitive environment, influenced by the integrated nature of the world trade and political systems. World war is unlikely due the looming threat of nuclear conflict, so instead political objectives are achieved through kinetic and non-kinetic proxy wars. Cyber and information are just some of the domains and environments where this competition plays out. Even from a warfighting perspective, in many cases the actual implementation of information operations as a component of military campaigning is focused on managing kinetic events against a narrative and coordinating non-kinetic effects to achieve specified outcomes. Therefore, the concept of information warfare can portray a more integrated and planned approach than is often the case. In reality, governments and societies, even totalitarian ones, must balance a variety of internal and external forces to shape their strategic objectives and narratives. For the purposes of this article, however, we will refer to information warfare in the context of ideologically and politically driven fake news, which seeks to manipulate, deceive, and change behavioral outcomes through disinformation for long-term strategic advantage.

While some aspects of campaigning fit within an informational economy, others bear a closer resemblance to the strategies and tactics of information warfare. Information by itself just informs an audience about the current state of affairs, leaving the parameters of decisions unchanged. Information becomes weaponized at the point that it shifts a target audience by either reshaping the environment or the preferences, attitudes, and even identities of the target audience in order to produce judgments, decisions, and behaviors favorable to the initiator (Marcellino et al., 2017, p. 9). This can be subtle. For example, rather than changing a person's desire to vote for a particular party, it is enough to simply convince someone not to vote on the day of the election. Value sets and beliefs are not always the target. Sometimes it is enough just to influence behavior for a short period to achieve the desired outcome.

The purpose of news is to inform the target audience which differs from the weaponization of information that seeks to deceive or manipulate through transformation of one's perception of a situation or through transformation of self-identity. This ties information operations to the rhetorical functions of communications of communication as information leaves things as they are, while rhetoric works on its subjects by influencing their identification with situations and their understanding of self-identity. The realignment of interests, attitudes, and beliefs through communications creates a "consubstantiality" between persons such that they come to see themselves as the same, at least within a certain set of parameters for acting (Burke, 1969, p. 21). Underlying the creation of consubstantiality involves shifting identifications with political objects and actors, as well as their understandings of themselves in relation to the political world. Identities are always relational, demarcating what one is and, simultaneously, what one is not (Connolly, 2002). Information warfare, therefore, involves the strategic and tactical use of information, which operates on the order of identities, shifting the alignments of a target from one set of political identifications to another, with the ultimate goal of shaping behavioral outcomes.

The focus on shaping identities and behavior is not limited to warfare between international adversaries. In contrast to the informational terrain of political conflict, which has informed models of spatial competition and political opinion formation, political preferences are not prior to campaigning but shaped by political identities, which are constructed through campaigning over time. Evidence points to the primacy of a politicized identity over information cues in understanding American political behavior. Political identity is irreducible to differences in issue positions as research shows policy positions even on fundamental issues such as abortion shift in line with partisan identities over time (Achen and Bartels, 2016; Mason, 2018). And while personality characteristics of candidates might be one calculation along with policy considerations, the explanation for political behavior predicated on the basis of an identity logic is quite distinct from the informational logic of policy preferences as policy preferences are derivative of partisan identifications rather than the other way around.[1]

An identity logic contrasts along all three dimensions of the informational logic. First, in terms of the target of definition, identities define actors rather than inform as to state of a situation in which action occurs. Personal identities are composed of "the commitments and identifications which provide the frame or horizon within which I can determine... what is good, or what ought to be done, or what I endorse or oppose" (Taylor, 1992, p. 27). The normative entailments of identities function in communications as an "inducement to action (or an attitude, attitude being an incipient act)" (Burke, 1969, p. 42).

Second, information and identity logics are temporally distinct. By overlooking the unique timestamps, deep fakes, computer-generated fake news, can work in deceiving targeted users. In contrast to the instantaneous and fleeting nature of information, identities temporally integrate an actor providing a sense of continuity over time and space (Miskimmon et al., 2014, p. 5). The repetition of identity claims perpetuates an identity narrative that preserves a sense of ontological security in the face of changing circumstances over time (Giddens, 1991, pp. 53–54). On the other hand, identities can be weaponized at the point that ontological security is put in jeopardy through communications that undermine one's trust in political actors and institutions or one's standing in the political system.

Third, in contrast to the descriptive nature of information, the moral horizons that define identities provides a language "for objects contain[ing] the emotional overtones which give us the cues as to how to act toward those objects" (Burke, 1984, p. 177). In online communications, "identity can be a shared feeling" as "people recognize themselves in the emotions of others" (Zaharna, 2018, p. 60). The contrasts to emotional appeals are descriptions of external objects and events without reference to the experience of those objects and events.[2] Emotional appeals can problematize identities as in the case of repeated communications seeking to induce anger or fear can give rise to anxieties—a tactic used by the Russian social media efforts to move Americans to deidentify with the existing political order (Jensen, 2018).

Fake News: Content Creation, Delivery, and Dissemination

Effective content creation relies on several different disciplines from target selection (military, political science, biology, psychology, and sociology) in service of creating memorable content. The content must resonate with the intended target, especially in an information-rich society. For this reason, a discussion of linguistics, psychology, and sociology is necessary. The delivery must be credible relying on psychology, sociology, linguistics, theater, and more recently data science. Dissemination often relies on technology, thereby introducing cybersecurity into the mix.

Linguistics: Analysis of Propaganda Tools

Linguistics is a discipline that is used in creation of deceptive data via rhetoric, but linguists are not consulted when countering efforts are necessary. Sample et al. (2018) cited the three well-known general attack types (ethos, pathos, and logos) as methods associated with supporting fake news and why these methods must be considered in any fake news countering solution. Of the three rhetorical groupings, each presents challenges as well as opportunities for automated processing by combining rules of linguistics and computer science when deploying computational linguistics.

The linguistic analysis of fake news must operate across a range of domains and disciplines, for every act of language is embedded in a context composed of a wide number of different influences and drivers (political, social, cultural, ideological... as well as the purely language-based). A study based on the identification of lexical and syntactic patterning as an identification and attribution tool (Conroy et al., 2015; Dey, 2018) can only be one step in the analysis. Similarly, a rhetorical analysis, detecting both individual figures of speech and appeals to ethos, logos, and pathos (see below), is vital, but can only be a single element in a much broader and deeper examination of the target texts. This article outlines a multidisciplinary approach for the identification and countering of fake news in general; we can also see how a blended, multidisciplinary methodology can operate at the linguistic and communicative levels. A helpful overview of one such blended approach is given by Zhou and Zafarani (2018), but in what follows, an outline model for the investigation of fake news is presented, based on analyzing it not as single acts of language, but as a *process* of targeted communication, consisting of several elements, all of which must be considered to permit a truly informed understanding of how fake news is *constructed*, and how it *functions*.

The following analytical schema is based on the seminal model of communication outlined by Shannon (1948), in which any communicative act is viewed as a process consisting of a *message* transmitted from a *sender* to a *receiver* via a *channel* (this is, of course, a simplification of the Shannon–Weaver model, but it is a useful starting point). The key thing to note here is that while we can study each element in isolation, a truly sophisticated analysis will consider how the various components

[1] This is not to say that identity-based logics cannot be subject to spatial modeling (see Glazer, 1995).

[2] This is not to invoke a strong subject–object distinction denoting the object of description need not define the content and sources of subjects.

interact and interrelate. For example, consider the issue of the chosen channel of transmission; different social media platforms operate in different ways (Twitter has restrictions of message length; Snapchat and Instagram are image-driven) and appeal to different demographics (Chen, 2020). Just as fake news must be carefully crafted to reach and appeal to specific target audiences, so any effective countering-strategy must consider the most appropriate communicative approaches and channels to mitigate against it. We know that there is a correlation between age of online users (and their political opinions) and their likelihood to retransmit fake news (Guess et al., 2019); work remains to be done on determining not just *why* this group is likely to fall prey to fake news, but in devising strategies for mitigating against this.

In order to conduct an effective analysis of fake news, we need to adopt the tools of corpus linguistics and establish a robust database (or *corpus*) of previous fake news campaigns. The examples cited in Tenove et al. (2018) and the two reports from the United Kingdom Information Commissioner's Office. (2018a,b) provide a helpful starting point. This will permit the creation of a detailed taxonomy of fake news, looking at sender/receiver/channel and allowing a detailed analysis of types of message and their specific linguistic/rhetorical features. While work has already been done in this field (Digital Shadows, 2018; Molina et al., 2019), there is a pressing need for a much larger set of corpora, which will permit a fine-grained analysis.

The final area for future research is that of traffic and social network analysis; in order to truly understand how fake news functions, we need to examine the ways in which stories spread across a platform, and how various users, such as trolls, bots, and super users, act as prolific spreaders of misinformation. The development of tools for collecting, analyzing, and visualizing message spread over time is a priority. The issue of timescale is vital; we need to consider, for example, what factors drive a particular item of fake news to disseminate rapidly across information space, while counter-messages often lag far behind and over a much smaller area. One tool that offers a useful basis for R&D across the entire domain of social media platforms is FireAnt, a piece of open-source software devised by Lawrence Anthony and Claire Hardaker (Anthony, 2018), and another freely available tool is the OSoMe tool developed at Indiana University (OSoMe.). This allows the capturing of data from Twitter over a set timescale; the data can then be visualized as a social network map, permitting the identification of key nodes of transmission. This in turn allows a fine-grained analysis of message spread and the possibility of targeting any countering-strategy toward the most prolific transmitters. The challenge will be to devise tools that can operate across the whole range of social media platforms (the image-driven, multimedia-rich domains of Instagram and Snapchat will be particularly testing).

Ethos: define a person or group

"Persuasion lies at the heart of political communication" (Flanagin and Metzger, 2014, p. 1); thus, the role of the messenger is highly relevant. For this reason, the messengers are targeted, as mentioned above. Gu et al. (2017) observed that the cost to discredit a reporter was $50,000. Flanagin and Metzger (2014) noted the role of credibility in presentation, as well as perceptions of honesty and fairness, even when the message remained constant. Ethos applied to the messenger will be further discussed in the archival data subsection of this article.

Ethos defines the target or messenger (Cockcroft and Cockcroft, 2005, pp. 28–54). The definitions for this activity can be positive or negative. In some cases, popular personalities or celebrities endorse politicians or politics (Scott, 2006) or cures, and in other cases, negative nicknames are used to both of these topics associated with a considerable amount of fake news. In some cases, trusted reporters are targeted in an effort to damage their credibility. A second type of ethos involves the messenger and will be further discussed in the archive section. Attacking the messenger can take the form of discrediting a reporter, a publisher, an editor, or any other entity in the news supply chain.

Fake news will inevitably build on the ideological/cultural/political values of the intended audience, or it will fail. Not only must it "speak the audience's language," as it were, but it must also operate within the frame of reference held by that audience. In so doing, the "us" against "them" narrative that is commonly deployed can take hold adding an emotional tie-in. This emotional tie-in is further discussed in the pathos subsection.

When ethos is deployed, the good–bad dichotomy prevails, even though most individuals have both good and bad personality traits, nicknames suggesting malevolence or benevolence with the target but not both. The use of ethos appeals to tribal identification and behaviors. Thus, this form of propaganda can be used with better efficacy on homogeneous groups in societies where fear of new or other groups predominates over curiosity or hope (Hofstede et al., 2010). In these societies, "in groups" are viewed as having significantly different values than "out groups."

Pathos: appeal to emotion

This rhetorical grouping is characterized by the appeal to emotions; thus, the emotions of fear (Montgomery, 2017) and hope (Menz, 2009) have a long-documented history of use in the political arena. Emotions play a critical role in political propaganda. In contrast to the descriptive nature of information, the moral horizons that define identities provide a language "for objects contain[ing] the emotional overtones, which give us the cues as to how to act toward those objects" (Burke, 1984, p. 177). In online communications, "identity can be a shared feeling" as "people recognize themselves in the emotions of others" (Zaharna, 2018, p. 60). The contrasts to emotional appeals are descriptions of external objects and events without reference to the experience of those objects and events.[3] Emotional appeals can problematize identities just as in the case of repeated communications seeking to induce anger or fear can give rise to anxieties—a tactic used by the Russian social media efforts to move Americans to deidentify with the existing political order (Jensen, 2018). Fear and hope have also been used in as motivators in military matters.

Of the three approaches, pathos is the most immediately effective, acting as a means of short-circuiting logic and rational thought and aiming to evoke an immediate emotional response.

[3] This is not to invoke a strong subject–object distinction denoting the object of description need not define the content and sources of subjects.

This is due to the emotional nature of decision-making. When appealing to pathos, punctuation can be a valuable tool. Punctuation was shown to act as a marker for propaganda in Israeli political discourse (Shukrun-Nagar, 2009).

Logos: appeal to logic

Any logos-based approach is highly challenging. Human beings are driven primarily by emotions, and the use of logical reasoning and data-driven persuasion will founder on the lack of statistical and general mathematical knowledge in the general public. Of course, statistics can be easily manipulated to customize narratives. In some cases, as noted by Pomerantsev and Weiss (2014), any narrative can be created, and a supporting reality can be created to support that narrative.

Partial truths and decontextualized facts. This type of logic is sometimes effective in political and scientific arenas. Consider the antivaccination movement that believed that vaccinations caused autism (Gross, 2009). An unexplained rise in the number of autism cases along with a discredited article from a scientist (Wakefield, 1999) with celebrity endorsements (Antrim, 2018) sounding credible created a combination of an ethos- and logos-based appeal. Many conspiracy theories contain elements of logos mixed with pathos. The messaging must include terms and resonate with the target audience; word choices should reflect similar or the same words that the targeted group would use.

Psychology: Understanding Individual Behaviors and Thoughts

Humans are fundamentally social creatures. Our social worlds are complex and require us to sift through information to determine what is truthful and what we need to know to maximize our survival and personal growth. However, the amount of information that we process is extensive and, in our digital age, delivered to us at high intensity. As noted previously, heuristics are strategies that we use to take cognitive shortcuts in order to handle this information overload (Tversky and Kahneman, 1983) and are also known as cognitive biases. On occasion, these biases may lead us to come to the wrong conclusion or to take the wrong action, but ultimately such biases are an adaptive strategy that helps us navigate our environment. Nevertheless, these biases may be targeted and exploited by the authors of fake news. For example, the secrecy heuristics may lead people to believe that any information that is presented as being from a secret source is more reliable (Travers et al., 2014). This heuristic is exploited in fake news stories that proclaim to reveal “leaked” information. The acceptance of fake news can be further increased through the use of images to accompany the narrative that is being put forward (Newman et al., 2012), which again may exploit biases by leading the reader to assume that a visual “record” is further evidence that a story is truthful. The inclusion of an image to accompany a fake news story also increases the likelihood that the story will be shared on social media (Fenn et al., 2019). Echo chambers (Boutyline and Willer, 2017) and filter bubbles (Holone, 2016) created through social media platforms may further reinforce these cognitive processes, through the aforementioned heuristic of confirmation bias (Kahneman and Tversky, 1973). Research suggests that the acceptance of fake news items can be combatted through the application of epistemic cognition, which refers to how we gather knowledge about our world and develop our belief systems.

Our understanding of our social world is also influenced by what we perceive to be the beliefs and attitudes of our peers. As noted above, this can lead to the proliferation of fake news, but the same social influence can also be a powerful tool in combatting fake news. Indeed, it has been observed that viewing critical comments from friends that challenge fake news items on social media is more effective, and prompting people to question the item themselves than a disclaimer from the social media provider stating that item appears to be fake (Colliander, 2019).

Emotions are another determinant of the acceptance or rejection of fake news, as predicted under the feelings-as-information theory (Schwarz, 2012). Falsehoods are 70% more likely to be retweeted than the truth (Vosoughi et al., 2018). This could relate to several other forms of cognitive bias. Survival information bias refers to our tendency to pay attention to information that relates to the health and well-being of ourselves or those we care about (Stubbersfield et al., 2015). An example of this would be the tales of poison found within Halloween candy that are shared among parents each year, despite there being no record of this ever happening (Snopes, 2020). Such stories that invoke survival information bias in turn prompt emotional reactions. Similarly, social information bias refers to our tendency to attend to information that represents some form of deviation from social values or social norms (Stubbersfield et al., 2015). An example of the exploitation of this type of bias would be fake news stories, which are based on a politician or celebrity being involved in a conspiracy of some type. This again stimulates an emotional response, which as predicted by the feelings-as-information theory (Schwarz, 2012) may be accepted as truthful. In keeping with theories of social gossip, this information is then shared throughout the individual’s social network, as it has been identified by the individual as being of importance (Mesoudi et al., 2006). This continual reposting and endorsement of a fake news item throughout a social network may contribute to the phenomenon of illusory truth, the effect exploited by marketers in advertisements for many years in which statements that are repeated are seen as being more truthful (Dechene et al., 2010), a phenomenon observed in relation to fake news stories such as those relating to vaccines and autism (Unkelbach et al., 2019).

One argument made is that people will tend to assume that information they are exposed to is truthful, as they draw upon their experience of the base rate where most of the facts they encounter in their lives are mundane and accurate (Brashier and Marsh, 2020). This assumption may be attributed to an anchoring bias where the target believes that because they are honest, news sources are also honest. This is compounded when the information is received from a trusted source (Flanagin and Metzger, 2014).

Differences in cognitive ability appear to be another factor that predicts how easily a fake news story can be countered, with individuals who have low cognitive ability being less likely to change their initial acceptance of fake information when explicitly presented with the correct information (De Keersmaecker and Roets, 2017). This also relates to another cognitive bias known as the anchoring heuristic in which people will tend to keep any subsequent judgments close to their initial judgment, even if that initial judgment is proven to them to be incorrect (Northcraft and Neale, 1987). A Cognitive Reflection Test is used by Pennycook and Rand (2019) to show that susceptibility to fake news is dependent thinking rather than partisan bias; others support this rationale by citing increased vulnerability because of reduced analytical thinking and less open-mindedness (Bronstein et al., 2019).

Sociology

Cultural values and divisions caused by cultural values are used to define and continue the dialog. In many of the most recent cases, these values are used to stoke divisions and amplify societal polarizations (Azzimonti and Fernandes, 2018). Interestingly, many of the divisions within the targeted society are tribal in nature, indicating that certain base values are similar in both groups (i.e., a desire to be treated equally; a desire to freely express one's views) but that the "in group" versus "out group" dynamic that defines tribes within a society is the targeted societal fissure. This is relevant because if the base values are the same, the targeted groups can be manipulated into thinking that the other group is the problem using many of the same techniques. An example of such behaviors may occur when one group within a society seeks equal rights, and another group perceives that in order for the first group to gain rights, they must lose or give up some of their own rights. While, logically this is not the case, the perception remains.

Higher technologic capabilities and interactions associate with greater vulnerability to information warfare campaigns (Szfranski, 1997). The information systems make possible the transmission of information at much greater speeds and volume, and interpretation of the information occurs with the human. Szfranski (1997) identified a strategic and tactical component that identifies identification (strategic) and restricting (tactical) disinformation. We argue that at the strategic level a fully integrated interdisciplinary response is required in order to accurately respond at the tactical level.

Orientation differs based on cultural values and heritage (Szfranski, 1997). Russia has been particularly active in stoking the immigration crisis in European Union countries (Volodymyr, 2016). By speaking to fears of "outsiders" invading countries, an increase in nationalism has arisen in host countries; specifically, Russia has created and supported right-wing narratives that speak to native citizens' fears of loss of cultural values and general well-being (ibid). Culturally speaking, most Western democracies exhibit a coexistence of high and low uncertainty avoidance (UAI) values (Hofstede et al., 2010). Hofstede et al. (2010) noted an association between nationalism and high UAI cultural values. According to Hofstede et al. (2010), high UAI values reflect a fear or uneasiness with the unknown, whereas low UAI values associate with a curiosity and willingness to learn about the unknown. The stoking of these and other cultural fissures or differences, in an open society, is easier with social media due to the wide reach of the communications medium.

Successful campaigns impose false realities on the human targets (Szfranski, 1997). Open societies are vulnerable to alternative viewpoints, the willingness to accept various viewpoints (Hofstede et al., 2010; Nisbett, 2010; Minkov, 2013; Sample, 2015), while normally a strength and defense against weaponized information (Szfranski, 1997), these cultural norms have been used against these societies through the promotion of carefully crafted false narratives that are in many cases quite sophisticated, and in all cases customized to the values of the targeted audience (Volodymyr, 2016; Sample et al., 2018).

Political Science: Influencing Policies

Fake news, or the tactical use of manipulative communications in a political context, serves to deceive political actors in relation to their strategic intentions regarding a situation. That may involve efforts to distort political positions and options so as to move citizens to vote (or refrain from voting) on grounds that misalign their preferences and actions. It may seek to cloud decision spaces for voters or political authorities by introducing spurious issues that misdirect attention. It may seek to inflame relationships between groups in a population by amplifying differences. In this sense, maligned fake news campaigns may be thought of as activities that place additional stress on political systems, thereby undermining their capacities to govern. On the other hand, a similar fake news campaign might deceive populations or political authorities in a direction that increases perceptions of trust, legitimacy, or the performance of the political authorities or the underlying system of governance (Easton, 1975). These are the elements of a political system that might be subject to support or stress; too much stress, Easton noted, could lead to system break down and violence). These latter efforts to support a political system by celebrating its achievements are a common practice in non-democratic countries such as the People's Republic of China where government employees are often called upon to promote the legitimacy of the CCP online, and an increasingly strict regime of censorship has prevented the emergence of critical commentary online (King et al., 2017; Jensen and Chen, forthcoming). Such efforts might artificially inflate the support for a system to serve the interests of an existing elite and order at the expense of efforts from the public to change the system. Either situation undermines democratic participation in a political system through distortions in the capacity to provide feedback and in its receipt by citizens and/or political authorities.

Political science treatments of "fake news" have focused on two distinct aspects. First, there is a question of whether foreign actors spreading fake news could have distorted the election results in 2016 in the United States. Second, scholars have studied the growing use of the term in recent time within political contexts, particularly its use as an epithet against journalists and perceived (other) political opponents. In terms of the interference question, the majority of the research suggests that it has had little impact on elections for two reasons. First, it is hard to distinguish Russian troll communications from other online sources, particularly those of the alt-right (Benkler et al., 2018). This delineation between Russia and alt-right publications is blurred since Russia today is a primary source of news stories for alt-right publications (Dorrell, 2017) that proceed to pass their article to mainstream conservative sites.

Related to this is the fact that domestic sources of news production and the campaigns themselves had considerably further reach than the Russian efforts, so it is unclear why Russian trolling would have produced an outsized effect compared to these other sources (Sides et al., 2018). Further, there is general doubt about the extent to which online manipulation campaigns have any effects based on research that shows political campaigning to general has neutral effects as competing campaigns cancel each other out (Kalla and Broockman, 2018).

Underlying the analysis in political science is a focus on information as the relevant unit of analysis in understanding the effects of fake news or manipulative campaigning. This focus owes its legacy to studies of voters and voter behavior, which have emphasized the role of campaigns and media outlets in transmitting information to voters who make up their minds (Downs, 1957; Ferejohn, 1990). Information efforts are considered independent of each other such that all information received, whether from a foreign state actor's manipulation campaign or domestic news sources, is equivalent in their potential effects. Jamieson (2018) notes that this is not necessarily the case as Russia has appeared to sequence its messaging in relation to information, which was not necessarily public at the time, and the sequencing of Russian activity and other actions can have unique and amplifying effects.

Further, although the literature in political science tends to find little evidence of a net effect of campaigning on voter choices, there are a few categories of places where they do find effects that may amplify the effects of targeted foreign interference activities. First, there is evidence that such campaigns have effects where candidates have controversial positions, and there is a lot of investment in mobilizing voters supporting those positions (Kalla and Broockman, 2018). That would be consistent with the normalization effects that a coordinated covert influence campaign can have by making certain positions appear more reasonable through repetition (Kahneman, 2011). Second, there is evidence that messages that provide grounds for people to express fear and to take a limited and discrete action based on that message (e.g., like, retweet, vote once) can help create a persuasive narrative identity narrative for a voter (Jamieson, 2018).

Beyond that, there is a literature in political science focused on "misinformation." Misinformation is often distinguished from disinformation in that the former is factually incorrect (a category overlapping with definitions of fake news), whereas the latter involves the intentional distribution of factually false claims for the purposes of inducing a political effect (Bennett and Livingston, 2018; Chadwick et al., 2018; Tucker et al., 2018). In relation to the misinformation literature, there is research into the efficacy of correctives (Nyhan and Reifler, 2010, 2015; Vraga and Bode, 2017). Some evidence suggests that false information can be corrected, but those effects tend to be limited to cases where an article of belief is not directly connected to one's belief and identity structure (Garrett et al., 2013).

Finally, there is an area of study in political science on the use of the term, "fake news," as a political epithet. Journalists and political opponents have been targeted with this term (Tandoc et al., 2018). Research shows that since the election of Donald Trump, there has been an upswing in the use of the term by politicians as an attack on others in places such as Australia, and the use of the term is usually amplified through reporting in mainstream news where it is not contested (Farhall et al., 2019). Propaganda historically has been understood by political scientists to not only provide a favorable narrative for one's own side but also to demoralize the enemy and undermine their will to continue the fight (Lasswell, 1927). There are wider corrosive effects on politics that some ascribe to this current era where the ability to know truths is often put into question, with some suggesting that we live in a "post-truth" political era (Keane, 2018). The consequence of undermining trust in political authorities, expertise, and expert systems can have many systemic implications beyond the discrete consequence of swaying an election as it can make a political system on the whole ungovernable through its polarizing effects (Singer and Brooking, 2018). Fake news today may involve a combination of foreign and self-inflicted wounds, which erode the will of citizens to participate in democratic political life—precisely the condition Tocqueville feared would give rise to a form of despotism (Tocqueville, 2010).

While the fake news has played a prominent role in politics, the movement of fake news into health and science is especially troubling. Particularly with COVID-19, the ramifications are serious, and in many cases deadly. In one case a couple chose to take chloroquine based on empty speculation about the drug and died (Neuman, 2020). In this case, the scientific community called for research, and the story as well as the scientific process became politicized.

Cybersecurity

Cybersecurity, a relatively new domain of war (Lynn, 2010), was slow to react to the fake news phenomenon despite the fact that information systems, particularly social media, were widely deployed in the targeting, delivery, and dissemination of disinformation. Some of the reluctance to engage centered around the censorship versus freedom of speech argument, but another reason for the reluctance was financial. Social media sites such as Facebook have business models that are heavily reliant on advertising money. Interestingly, Taboola has been a common advertiser associated with fake news (Neilsen and Graves, 2017).

The argument continued, but a change occurred when the role of Facebook and other social media sites became known in the selling of personal information to Cambridge Analytica (Cadwalladr and Graham-Harrison, 2018; Risso, 2018), where the data were mined for use in political circles. The sharing of personal information crossed the privacy boundary that is one of the core tenant areas of cybersecurity (McCumber, 1991). The sale of user personal data (Cadwalladr and Graham-Harrison, 2018; Risso, 2018) and the customization of disinformation for the targeted users supported Szfranski (1997) assertion that civilians as well as military personnel will need protection from disinformation.

Szfranski (1997) observed that information systems are the primary means by which adversaries collect information on targets. Artificial intelligence (AI) algorithms that create and re-enforce filter bubbles (Sîrbu et al., 2019) can amplify the development of echo chambers, reducing the individual's ability to find unbiased news. Furthermore, personas in the filter bubble in agreement with the target's views gain credibility with that group. Filter bubbles are further discussed in *Data Science: Processing Large Volumes of Fake News* of this article.

The speed and reach of the internet enabled the rapid spread of disinformation, suggesting that any viable solution to counter the spread will require the automation and detection associated with cybersecurity to play a role in this endeavor (Cybenko et al., 2002; Horne and Adali, 2017; Sample et al., 2018). Furthermore, Russian campaigns have been characterized by the intensity of their use of information systems to promote narratives (Volodymyr, 2016; Payne, 2017; Jensen and Sample, 2019).

Volodymyr (2016) identified large-scale ongoing hybrid operations that relied heavily on Internet connectivity and social media outlet including Syria (Turkey), the European Union, and Ukraine. Since then NATO and liberal Western democracies have been added to the list for operations that originate in Russia. Other countries and groups have been encouraged by Russia's success. The global nature of the problem that resulted in an unanticipated use of Internet technologies made digital deceptions such as fake news and deep fakes a problem that now belongs in part to cybersecurity.

The drivers for change within the cybersecurity field are not currently strong enough to lead to the development of solutions. Until nation states seek to apply larger-scale solutions and adopt policy changes that encourage the security of individual privacy rights, the cybersecurity industry has no incentive as a financially driven business area to address the problem. Indeed, commercial bodies have helped create this situation by repeatedly removing individual rights for privacy on the internet and creating a new norm, accepted by users, which they should trade their data for convenient services. A combination of commercial and legal drivers with strong law enforcement and government agency support has eroded anonymity on the internet. A strong incentive to change existing norms is likely to come from the application of targeted information operations on political, military, and critical national security personnel on a scale that forces policy changes. However, without such a driver, until this paradigm changes and there are legal or financial costs associated with the commercialization of user data on the Internet, the main drivers enabling online fake news will continue. Unfortunately, the commercialization and lobbying that have emerged in this area make it unlikely that the cybersecurity industry will be able to address this problem in the near term, except as a supporting role to respond to policy change. Efforts to integrate financial payment systems with social media and news outlets are likely to further exacerbate the problem. In the meantime, greater education of users and policy makers and focused attempts to secure users through anonymized and secure applications are probably the best short-term solutions available to cybersecurity practitioners.

Data Science: Processing Large Volumes of Fake News

The work performed by Cambridge Analytica (Risso, 2018) exemplifies the growth of data science in the fake news space. Originally used in marketing, another discipline that processes inputs from psychology, sociology, and the arts, the algorithms have been more recently applied to political goals. There are several different types of algorithms ranging from history-based heuristics, through trees and neural networks. Each of these algorithms can be manipulated through their data to either re-enforce preferences or to steer preferences into a new direction. Furthermore, the algorithm-created filter bubbles that re-enforce beliefs are amoral and probabilistic in nature; thus, the stories presented to the user are similar in tone and accuracy.

The ability to manipulate AI outputs extends beyond fake news propagation and into all aspects of AI and machine learning (ML). The example of Tay (Risley, 2016), the Microsoft chatbot that was trained to be helpful, but ultimately became abusive, illustrates the ability to manipulate AI and unintended direction through input manipulation of legitimate data. Weight and data manipulation fuels research into adversarial machine learning (AML) and malicious use of AI (MUAI), and many of the deceptions that other domains encountered occurred earlier in fake news.

The application of AI and ML in the creation, dissemination, and countering of fake news represents a growth area in data science, and as such, much has yet to be discovered. Deep fakes provide an example of this growth area. While data science can provide fascinating new insights to existing problems, the examples listed above remind us that the power of data science techniques can be used for benign, malevolent, or benevolent applications. A key point to remember in data science is the importance of the query being processed. The phrasing of a query sets the algorithm on a weighted path based on the data learned during the ML process. Depending on the data classifications, results can vary to the same query; this action is observable when two people enter the same query into a search engine and receive different results.

Despite the problems that AI and ML introduce, combatting fake news and other digital deceptions will be a job for AI/ML; Horne and Adali (2017) have already shown successes in using ML trained hosts to detect fake news. The accuracy will improve as the rules become more complete. Should the rules lack completeness, then fake news detection will be limited to detecting stories that fit a known pattern, signatures in cybersecurity parlance, and signatures have a low to non-existent detection rate with novel approaches. AI/ML are extremely proficient at detecting patterns, but the patterns must be familiar to the software, hence the need to create rules based on knowledge garnered from other disciplines.

Theater: News as Entertainment

"All the world is a stage" (Shakespeare, 2009), and fake news has a theatrical aspect. Theater can be thought of as a communal art form where the audience and performer share the roles of subject, spectator, and benefactor (Abaka, 2014). Similarly, fake news, particularly in an interactive mode of delivery (i.e., talk radio, live interactive TV news shows, social media, etc.), creates the same dynamic.

Theater, like fake news, is also interdisciplinary. The theater consists of eight major disciplines:

acting, directing, writing, producing, costumes, set, sound, and lighting design (Jones, 2004). Each of these disciplines is discussed along with their role in the creation and delivery of fake news in support of seeking a shared, desired intention within the story and the audience.

Theater requires the disciplines working together to create or manipulate the audiences' thoughts and feelings. Fake news particularly uses propaganda delivered on trusted mediums in personal spaces such as talk radio in a car, TV in the home, and social media on cell phones and personal computers. All of these devices are trusted transmission sources for the user. The messaging deploying the use of previously described propaganda techniques is delivered in an environment that the user has already deemed trustworthy.

The story is most effectively told when the disciplines fuse together, creating a final product that is greater than the sum of the collective parts Jones (Hostetter and Hostetter, 2011). Actors use voice, facial expression, and body movement to convey the message that captures and unites the conscious and unconscious minds through shared presentation and decoding of words and symbols (Hostetter and Hostetter, 2011). Similarly, reporters and commentators use their voice, facial expressions, and body movement to both consciously and unconsciously deliver a message.

Actors prepare by not only memorizing lines, but also by drawing on emotional experiences that audiences can easily understand and find relatable, by using physical movement through expressions and body movement to convey emotions bringing to life the text of pathos. Once the message has been crafted, the delivery must seal the emotional hold.

Once the message has been successfully delivered and sealed, the re-enforcement can be taken care of through secondary actors known as trolls and bots. Trolls, paid personnel used to amplify a message can assume the role of unseen actors in support of maintain the target's engagement. Bots are the automated counterpart of the troll, performing the same duties through the use of AI.

An actor seeks to tap into the audiences' emotional memory (Stanislavski, 1989). Sight and sound are entryways into the imagination. Once an actor commands these two senses, they have a pathway into the imagination and can direct and manipulate the mind. Stanislavski (1989) considered sight and sound the two primary senses, and touch, smell, and taste secondary senses that can be triggered through the primary senses. Stanislavski (1989) noted that if an actor can appeal to only one of the senses, the remaining senses will also be available to the actor, all in support of influencing the audience's emotional memory. One part of the director's responsibility is to work with the actors to set the tone and intention for the play. The news director works with the news anchor to convey the tone and content of the story. The same news story compared across news stations can vary in length, tone, and detail (Sample et al., 2018). In the case of interactive news, much like in the theater, mood and emotions can be extracted from an audience.

Another important role that the theater director plays is in deciding what parts of the story remain and what parts are cut. Similarly, in the news media, the director/producer/editor determines which stories, or portions of the selected story, are presented to the audience. Recalling the earlier discussion on thinking, an event that is not sensed is not perceived. As with AI/ML, the flow of information is controlled by an entity in support of a specific intention.

The writer relies on words, language, rhythm, pacing, and musicality of language to be delivered by a skilled actor or anchor. All are well-crafted and integrated to achieve the intention of the story. The integrated whole is executed to play upon the values and beliefs of their audience. The use of memorable quotes or rhymes creates an indelible memory for the audience and the actor. Word choice for the story writer fits into manipulative linguistic choices, particularly those associated with pathos. Easily remembered phrases associated with fake news are effective methods of delivering a memorable message.

The producer is the person in charge of operations including hiring of directors, actors, and other personnel who support the vision. Casting of anchors and reports sends an enormous conscious and unconscious message to the audience. Attractive speakers benefit from the halo effect, while less attractive speakers are perceived as being more untrustworthy (Zebrowitz and Franklin, 2014). This finding suggests that a less attractive person would need to compensate for the untrustworthy impression, possibly by using intellect. The producer is responsible for all aspects of the production from vision, through operations and budgeting. In short, the producer is the production CEO possessing the holistic vision.

Costumes add a sensual authenticity to the production. Something as seemingly small as hair and make-up can have a significant effect on how the audience perceives the actor in theater or the reporter for news. Consider the appearances of various news hosts and the consistency of appearance on each of the major news stations. Suits for men are the costume with ties being carefully chosen. Women, while not forced to wear suits, also have clothing rules that allow for some flexibility of choice (Hillman, 2013). Within the range of professional clothing attire, a range exists for various articles of clothing and accessories (Crilli, 2014; Moeslein, 2019). A clear difference can be observed when viewing before and after pictures of liberal and conservative program hosts.

Set design provides the context or the visual environment. Color can set the emotional tone. For news, a simple calming blue background attracts attention, as does the color red (Hillyard and Münte, 1984), which suggests serious fact-based messaging. Red is a universal alerting color that people are trained to stop and focus their attention on (Kuniecki et al., 2015). Breaking news alerts appear on a red background. When delivering fake news, a set matches the color scheme and set design of the traditional mainstream media news station where the news anchors are seated but leaning forward. A well-thought-out set design seamlessly supports the reporter's performance.

Audio is the presence or absence of background sounds such as music; voices and other sounds are also used to set the mood and manipulate emotions. Dark low-pitched music accompanied with dark lighting and a dark background suggests a sense of foreboding, preparing the audience for bad news.

Lighting is the remaining theatrical discipline. Lights and the shadows cast by lights allow for creating a perception. Shadows cast over an object suggest something to hide or untrustworthiness. Bright light suggests honesty and integrity. A harsh bright light suggests shining a light on a dark or dirty subject, exposing what was hidden.

All of these to point out the conscious effort of an entire production team to seek out the desired behavior, thinking, and feeling of a target audience. When executed with precision and art, the audience has no real defensive against it. Even the best of professionals in this field find themselves taken away and moved by the production. The actor must live it, feel it, and experience the depth of human nature supported by all the other disciplines to achieve the goal of a super objective set by producers, directors, and writers. How better to manipulate people into seeing things our way and buying in to joining our team.

For years, the delineation between news and entertainment has continued to blur (Edgerly, 2017). News reporters, while publicly claiming neutrality, can convey messaging through expressions, intonation, and movements, all methods of non-verbal communication that actors draw on. Reporters can convey joy, anger disgust, outrage, and other emotions without changing the text of the story. The goal of all news reporters, much like an actor, is to keep the audience's attention. To that end, reporters, particularly television and other video reporters, like actors, wear make-up, perform in well-lit conditions, and engage with the camera. Through costume, make-up, set design, writing, direction, lighting, and sound, the reporter, much like the actor, sets out to elicit the audience's emotional and physical response.

Cybenko et al. (2002) noted the mundane aspects of factual data. While linguistics offers a strong starting point and works well with textual data, the staging of events such as protests (Gu et al., 2017; Mueller, 2019, p. 29) requires inputs from theater. Theater, like all art, relies on sensory stimulation. Theater integrates sensory experiences relying on triggering sight and sound directly and touch, smell, and taste indirectly. These inputs are designed to trigger sensing from the target, to shape the perception.

LESSONS LEARNED FOR COUNTERING FAKE NEWS

Because the problem is interdisciplinary in nature, any response that does not take into account the various disciplines can only partially succeed at best. This means that models must be superimposed and incorporated into the response. While an automated response favors data science and cybersecurity, this response will not fully succeed if linguistics, psychology, and sociology, along with the frameworks defined in those disciplines, are not fused into the solution.

Data science provides insights but requires carefully worded questions and subject knowledge in order to form the best, most comprehensive queries. Cybersecurity-informed solutions tend to be reactive, suggesting that queries posed by cybersecurity personnel may lack abstraction and will result in no new insights even when new problem sites are discovered. Furthermore, the overall reactive posture may result in software missing new styles and techniques in fake news creation and dissemination.

When combining data science and linguistics, computational linguistics offers some of the tools that can enable rapid detection of propaganda; however, natural language processors (NLPs) have shortcomings that may be unfamiliar to data scientists. NLPs used in computational linguistics are able to quickly synthesize large volumes of data presenting common themes through "bag of words" outputs and sentiment analysis. An explanation of problems in each area follows.

NLP Problems

Natural language processors group together common words and group emergent patterns in written text. In order to do so efficiently, several things must happen that can result in misleading outputs for the data scientist who programs the software, including punctuation removal, case changing, word-stemming, and filler word removal.

- Punctuation removal is problematic because "!" and "?" and quotation marks are all a part of the sensationalism that elicits emotional responses from the reader (Shukrun-Nagar, 2009; Cohen et al., 2018).
- Case changing occurs when software evaluates words the ASCII representation for "News" is different than the one for "news." In order to work around this problem, uppercase letters are converted to lowercase, and in most cases, no problems result. However, with emotion-driven fake news, words in all uppercase are present to stimulate the visual sense; by changing these words to all lowercase, an important textual clue is removed.
- Word-stemming results when NLPs convert adverbs to verbs, or convert verb tenses to the root word, so that "quickly" and "quicken" become "quick." The problem is adverbs are words that are used to elicit an emotional response (pathos); by stemming these words, another textual clue is removed.
- Filler word removal is performed when NLPs remove words such as "the," "and," and "he" with the goal of avoiding filler words outweighing nouns and verbs. The problem is that the removal of pronouns also results in the removal of "us," "we," "they," "them," important words in the "us" versus "them" narrative oftentimes deployed in propaganda.

The problems listed above are easily solved when the programmer is aware of their existence, and most often, the programmer remains unaware. The problem is that the NLP packages are a part of a larger software development effort, and the aforementioned examples are not considered. The developers do not know what they do not know. In many cases, packages can be modified, or preprocessers can be written that quickly address these problems. Much of NLP work aids AI/ML rules, so knowing the correct query to make falls outside of the expertise of the programmer tasked with writing the software.

Countering Content

Cybenko et al. (2002) noted the importance of detecting the attack before the narrative can affect the target's behavior. Of course, this suggestion seems to work posteriori as a method to prevent repeat mistakes. One suggestion was to detect the preconditions that exist as a method to inoculate the target (ibid). Another would be to incorporate rules of propaganda into computational linguistics. The authors propose that three-point model of computational linguistics to address contextualization of data, as well as the descriptiveness, pattern-spread analysis to address the temporal aspect, and archival reputation analysis that adds temporal, context, and descriptive values for additional analysis.

Computational Linguistics—Contextualization and Descriptive Analysis

Fake news, by its very nature, consists of texts. Sometimes multimodal (employing, for example, audiovisual material), but overwhelmingly written compositions, designed to be transmitted (and retransmitted) across electronic/digital media to a carefully selected target audience. As communicative artifacts, they are open to analysis through linguistic tools, and as persuasive communications, they can be related to the body of work that has been done over centuries relating to rhetoric (the art of persuasive communication) in general, and propaganda in particular (see, *inter alia*, Ellul, 1965; Jowett and O'Donnell, 1986; Pratkanis and Aronson, 1992; Connelly and Welch, 2003).

Despite the list of problems enumerated above with NLPs, for those engaged in combatting fake news, the ability to employ computer-aided social network analysis offers great potential for mapping (and ideally countering) the spread on misinformation online, by quickly identifying the key vectors of fake news and tracking (in near-real-time) the flow of fake news across the Web (Hardaker and McGlashan, 2016; Chetty and Alathur, 2018). The use of Computational and Computer-Aided Corpus Linguistics (the identification of key textual features through comparison of a target text with a corpus or corpora of reference texts) offers a real possibility to create automated tools for the identification of fake news through its linguistic content and its removal from social media without human intervention and for automated generation of effective counter-texts (Pérez-Rosas et al., 2018; Marquardt, 2019; Pathak and Srihari, 2019).

The rules of propaganda are well understood, and the ability to modify software to enforce rules is attainable, particularly when ML is combined with statistical tools such as group comparisons and correlations to compare text. Many of the lexical expressions found in propaganda are synonyms, which can easily translate to neighboring words in ML classifiers (e.g., "step" and "leap"); thus, when phrases containing neighboring words are evaluated computationally, they will show that the distance between is statistically insignificant. Compared against the traditional phrases used in conversation where the word distance is statistically different, the ability to analyze and provide decision-making assistance is near real time.

The other areas listed above can be quickly evaluated using simple linguistic tools or even scripts that can strip out and quantify punctuation symbols before they are removed without record. Similarly, the stemming problem can also be addressed as a part of the preprocessing while preserving the important metadata. Computational linguistics works alongside data science and can provide much of the inputs needed to feed training data rules.

Pattern Spread—Temporal Analysis

In order for fake news to spread, stories must be artificially promoted using trolls and bots (Rosenblatt, 2018). The use of these aids leaves digital traces. Vosoughi et al. (2018) illustrated this phenomenon showing the extreme volume of stories that saturate the news cycle. When combined with the credibility of receiving these false narratives from trusted sources and the speed in which these stories spread, preemptive inoculation can become problematic. The OSoMe toolbox developed at Indiana University[4] is another tool that illustrates the overwhelming spread of fake news. So that even the informed reader, when trying to gather information on alternative views, not only does not receive those stories due to AI algorithm weighting, but also the overwhelming volume of stories with the fake narrative (Gu et al., 2017; Sample et al., 2018; Jensen and Sample, 2019). Presently fake news spread patterns are rather distinct, but since they rely on automated software. they can be easily adjusted. In addition to providing a dominant narrative in volume, this flooding behavior also manipulates sentiment analysis, providing deceptive data to analysts.

Suggestions by Cybenko et al. (2002) included information trajectory modeling as a countering tactic. Information trajectory allows for comparisons against historic data and distance measuring from the historic data. This approach can be used to model pattern spread as well as linguistic differences and was proposed by Sample et al. (2018). However, unlike the linguistics component, there exists a lack of data for pattern spread of factual narratives, and factual narratives will have varying baselines depending on the nature of the story. For example, a natural disaster with many casualties will show a different pattern spread than a special interest story, which differs from a news story surrounding a celebrity. If the fake pattern is the area of focus, the parties responsible will alter the behavior to keep the filters from detecting the pattern; in cybersecurity, this is known as "fuzzing." "Fuzzing" occurs when a character is changed in the signature string, allowing the new malicious data to slip in, undetected by the security filters.

Metadata Analysis of Archival Information—Time, Content, Context, and Reputations

Data science has been used and suggested for solutions to digital deception. This shows the enduring value of messenger credibility as discussed by Flanagin and Metzger (2014). Reputations can be discredited for an affordable cost, and reporters are human,

[4] http://osome.iuni.inu.edu/

so mistakes will happen. A gullible reporter who is repeatedly fooled may not have a very long career in his/her chosen field, but a reporter who consistently reports the facts, with or without theatrics, will have a pattern that is worthy of being considered credible. Currently, this process is performed by humans when they evaluate the credibility of a source, and as we previously discussed, the human decision-making process in this area is flawed and under attack.

Data science can go beyond sentiment analysis and some of the other techniques discussed earlier in this article. Every proposed countering technique will generate metadata, data about the collected data. This data are prime for fresh insights. Some of the metadata fields of interest would include, but not be limited to, reporter information, publisher information, time information, context content, and linguistic characteristics. A brief discussion follows.

- Reporter information: Reporter bias scores, average story word count, average linguistic characteristics associated with reporter including % nouns, verbs, adjectives and adverbs, publisher associations, credibility score, and sources used.
- Publisher information: Publication bias scores, average story word count, average linguistic characteristics, reporters used, credibility score, average story lifecycle, average time to report, sources used.

Time information is important because fake news typically has short lifecycles because once the deception is discovered, not everyone wishes to propagate it further; thus, time distance from original source, lifecycle, early source trajectory.

This archive of metadata with the small sample set of fields provides a starting point and is by no means complete. This starting point allows for groupings to answer known questions and unstructured analysis groupings to inspire new lines of thought and question. Furthermore, by collecting and processing metadata, the actual use space is smaller, resulting in a smaller and more efficient archive, which contains links to the larger complete archives.

The Role of Artificial Intelligence and Machine Learning

The need to understand the rules in order to create AI/ML rules otherwise signatures on steroids is discussed. Using ML software and satire as training data, Horne and Adali (2017) were successful in demonstrating software could successfully detect fake news in the form of satire. While the findings were encouraging the study looked for a specific type of fake news, satire.

AI is highly dependent on the classification schemes attained through ML, and both AI and ML are highly dependent on the accuracy and veracity of the training data used. The fact that in cybersecurity the poisoning of training data is an area of research should serve as a cautionary point. When rules are done well, AI and ML can easily outperform humans on many tasks; however, questions remain open on AI biases that may be intentionally or unintentionally inserted by the programmer (Bellovin, 2019). Biases have been observed in facial recognition software (Nagpal et al., 2019), resulting in unanticipated outputs. The importance of balancing the training data requirements results in having to balance inadequate data that yield false results, or overfitting provides accurate findings but with little to no abstraction. Striking this necessary balance requires an understanding of rules that lie outside the discipline of data science and cybersecurity, the two disciplines in the enforcement arena, and into the behavioral sciences space.

Game Theory to Inoculate—An Alternative Approach

Immunizations against diseases occur from patient exposure to a weakened form of the virus; this forms the basis for work done at Cambridge University (Van der Linden et al., 2017; Roozenbeek and Van Der Linden, 2019). While game theory was not discussed above, game theory does incorporate many of the aspects discussed in the previously mentioned disciplines that feed decision-making. The researchers found that by exposing participants to fake news stories. The researchers at Cambridge introduced the topic of attitudinal inoculation where subjects are warned preemptively, and false narratives are preemptively debunked, so that when exposed to deceptive data, they discard the information. Similar approaches work in cybersecurity when users are preemptively warned about spam and malware (Bindra, 2010). The research shows promise but requires prior knowledge of the false narrative before the narrative is created and distributed. In some controversial areas such as politics and climate, science may be difficult to anticipate the new narrative, or in cybersecurity terms, the zero-day narrative. In other areas, such as fake news surrounding medicine or economics news, the scope is more limited, making possible success in the exercise in anticipation.

EXAMPLES USING THE MODEL

The three-point model relies on evaluating using linguistic features, pattern spread, and archival reputation analysis. Included are some examples of the evaluations. On February 16, 2020, when the COVID-19 pandemic was in the early stages, Dr. Anthony Fauci was interviewed by CBS news on the show Face the Nation. A transcript of the show[5] can be found in the public domain. Dr. Fauci's remarks about the virus were approximately 646 words, containing no special punctuation, and the ratio of adverbs to text was 1:215. A few days later on February 25, 2020, Rush Limbaugh also shared thoughts on the same subject, the virus, during his talk show. Mr. Limbaugh's text was also made publicly available[6]. Mr. Limbaugh's word count was 756 words, containing six special punctuation instances, five "?" and one "!" along with an adverb-to-text ratio of 1:126. Additionally, Mr. Limbaugh's text invokes rhetorical devices discussed earlier such as plain folk talk. Thus, based on these few characteristics, Limbaugh's words would deviate farther from ground truth than would Fauci's. This is not to say that Dr. Fauci's remarks are considered ground truth, but that Dr. Fauci's

[5] http://cbsnews.com/

[6] http://mediamatters.org/

words are measurably more trustworthy than Mr. Limbaugh's. Limbaugh's text would suggest more than one standard deviation off of ground truth, and Fauci's remarks would be considered less than one-half standard deviation from ground truth.

This brings us to the pattern spread. The OSoMe tool[7] allows for rapid tracking and visualization of hashtags and other social media features. When the Fauci and Limbaugh hashtags are compared in **Figure 1**, Limbaugh outperforms Fauci. Rush Limbaugh had been awarded the medal of Freedom in early February (notice the peak in the Limbaugh hashtag, followed by a drop that remains higher than Fauci's even when Fauci's interview is aired. Even more interesting is the delta between Limbaugh and Fauci on February 16, the date that the Fauci interview aired. The pattern spread appears to show the fact-based narrative underperforming. Only in March, once the pandemic had taken hold did Fauci surpass Limbaugh. Limbaugh had a full month (possibly longer) to deliver his version of the message to a larger audience.

This same phenomenon can be observed again with the face mask debate where "#nomask" drowns out "stopthespread,"

[7] http://osome.iuni.iu.edu/

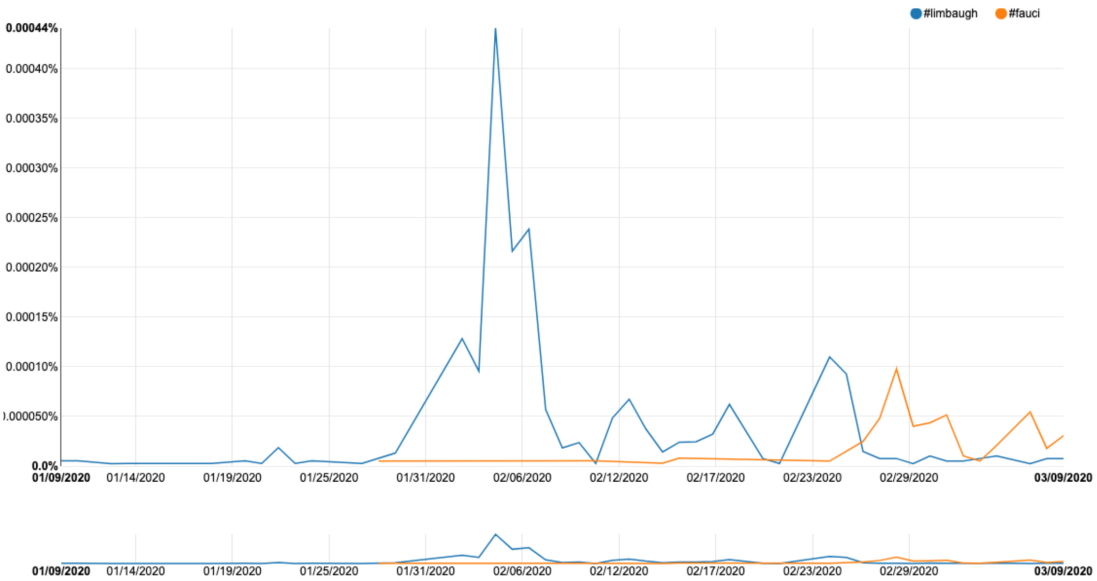


FIGURE 1 | Trending data Limbaugh and Fauci mid-February.

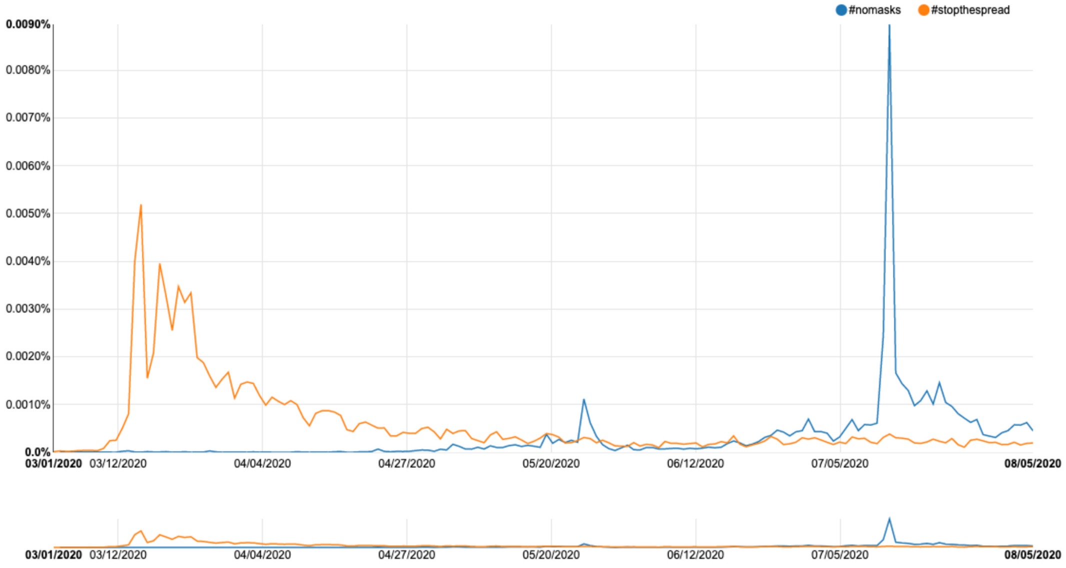


FIGURE 2 | Hashtags from the mask debate.

illustrated in **Figure 2**. Then more recently, this pattern is observed again with the "#movetheeletion" hashtag when compared to the real news of the sharp decline in GDP in the United States, as seen in **Figure 3**. The underperforming hashtags typically distribute longer, and the peaks tend to be shorter than their counterparts associated with fake narratives. Thus, the fake hashtags would result in the deviation value increasing. In the case of the #nomask hashtag, this pattern of a steep, dramatic peak with small tails was observed with election data in 2016 (Sample et al., 2018).

False or misleading narratives are also used to distract from factual news stories. This example was recently observed when President Trump suggested delaying the November election as the GDP quarterly data were released. In this case, the large drop in the US GDP would normally be the leading news as this drop would be associated with a significant economic recession. However, **Figure 3** illustrates the Twitter feed of tweets for delaying the election that drowns out the GDP data tweets on July 30–31, 2020.

Pattern spread data can be valued by measurement off of known good news stories pattern spread.

The final model point of reputation analysis can be thought of as a consideration of the source. In the Fauci–Limbaugh example, the source of the Fauci interview is widely considered reputable. In this very simple example, Dr. Fauci has a history of speaking accurately, particularly when discussing medical matters; thus, his overall accuracy value would be considerably better than that of Mr. Limbaugh. This reputation analysis could be further enhanced by examining the sources that picked up both speeches for publication and those sources could also be evaluated. A final scoring metric could come from fact checking organization, where a point is also given for debunking a story narrative. The overall analysis when the three-point model is applied is that the closer to zero the total score (linguistic, pattern spread, and reputation values) lands, the more factually accurate the story.

SUMMARY

Szfranski (1997) argued that information warfare attacks would be enacted against both information systems and belief systems and that leaders and their supporting non-combatants would both be targeted. This is currently the case with fake news. Szfranski (1997) thought that open societies, such as Western democracies, would have better defenses than their autocratic counterparts, but the results on this front are decidedly mixed. The current iteration of fake news or propaganda operates in a jujitsu fashion, where a target's strengths are used against itself, something that was unanticipated with the rise of digital propaganda.

There are many different reasons for the widespread success of this iteration. Key factors include the refined targeting techniques, the breadth of the Internet reach, the trust of social media platforms, the financial incentives for data management companies to sell information, and the use of open values systems found in open societies against those societies. Some open societies (i.e., Finland, Estonia, Latvia) have demonstrated resilience to fake news (Atkinson, 2018), suggesting a possible common set of values that have not been investigated. However, implementing defenses or countering tactics at speed requires an interdisciplinary approach. Meeting this challenge requires deep interactions that reflect a true exchange of ideas and

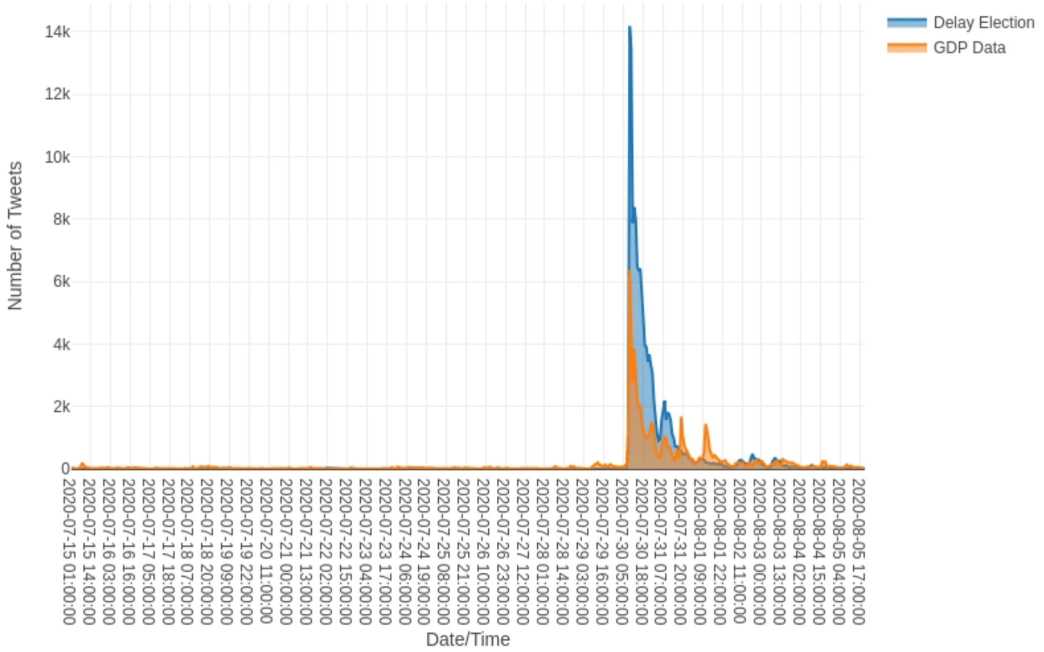


FIGURE 3 | Twitter feeds for GDP drop and delay the election.

implementation of models outside the traditional disciplines that support a fused response to this new generation of espionage (Younger, 2018).

Interdisciplinary work can help provide a more unified front to a currently fragmented, open-ended institution (Waisbord, 2018). Cybersecurity can aid in explaining how social bots spread misinformation; this can help journalists understand how misinformation spreads online and address increasing frustrations controlling fake news propagation (Schapals, 2018). Psychology sheds light on the phenomena of fake news amplification by means of echo chambers and confirmation bias to provide more conscious perspectives of how news, whether fake or not, is consumed in different cultures and groups. Akin to other areas, for journalism to adapt to change cultivated by the internet, a discussion should be initiated that recognizes the integration of multidimensional solutions. Journalists not only are responsible for the integrity and truth of their own reports, but also recognize that existing understanding of fake news spread is foggy (Mhamdi, 2016). Interdisciplinary approaches should acquit journalists from being solely responsible for policing fake news in today's contemporary digital environment and emphasize collaboration to account for other trends responsible for fake news proliferation in our information streams.

AUTHOR CONTRIBUTIONS

CS: lead author, contributed to all sections, primary authorship on sociology, cybersecurity, data science, sections "Lessons Learned for Countering Fake News" and "Examples Using the Model." MJ: sections "Introduction" and "Background," primary authorship political science. KS: sections "Introduction" and "Background," primary authorship linguistics. JM and AB: primary authors psychology. SF: primary author theater arts. DO and AO: secondary authors deception and cybersecurity. All authors contributed to the article and approved the submitted version.

FUNDING

JM and AB are in receipt of funding from the US Army Research Laboratory for a match funded Ph.D. studentship entitled 'Fake news, culture and heuristics,' grant number 11708.

ACKNOWLEDGMENTS

The authors wish to acknowledge Professor Sin Ming Loo from Boise State University for his support in reviewing this effort.

REFERENCES

Abaka, A. (2014). *Audience Development as a Communal Experience*. Available online at: https://howlround.com/audience-development-communal-experience (accessed on March, 19 2014).

Achen, C. H., and Bartels, L. M. (2016). *Democracy for Realists: Why Elections Do Not Produce Responsive Government*. Princeton: Princeton University Press.

Anthony, L. (2018). Introducing Fireant: A Freeware, Multiplatform Social Media Data-Analysis Tool. *IEEE Trans. Pro. Commun.* 61, 1–15. doi: 10.1109/TPC.2018.2870681

Antrim, A. (2018). *Anti-Vaccine celebrities have inordinate amount of influence. Pharmacy Times*. Available online at: https://www.pharmacytimes.com/resource-centers/immunization/antivaccine-celebrities-have-inordinate-amount-of-influence (accessed on June, 18 2018).

Applebaum, A., and Lucas, E. (2016). *The Danger of Russian Disinformation*. Washington, DC: The Washington Post.

Armistead, L. (2004). *Information Operations: Warfare and the Hard Reality of Soft Power*. Sterling, VA: Potomac Books, Inc.

Atkinson, C. (2018). Hybrid warfare and societal resilience implications for democratic governance. *Inform Sec.* 39, 63–76. doi: 10.11610/isij.3906

Azzimonti, M., and Fernandes, M. (2018). *Social Media Networks, Fake News, and Polarization* (No. w24462). Cambridge: National Bureau of Economic Research.

Bakir, V., and McStay, A. (2018). Fake News and The Economy of Emotions. *Digit. J.* 6, 154–175. doi: 10.1080/21670811.2017.1345645

Bellovin, S. (2019). *Yes,'algorithms' can be biased. Here's why. ARS TECHNICA*. Available online at: https://arstechnica.com/tech-policy/2019/01/yes-algorithms-can-be-biased-heres-why/ (accessed on Jan, 24 2019).

Benkler, Y., Faris, R., and Roberts, H. (2018). *Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics*. New York, NY: Oxford University Press.

Bennett, W. L. (2012). The personalization of politics: Political identity, social media, and changing patterns of participation. *Ann. Am. Acad. Politic. Soc. Sci.* 644, 20–39. doi: 10.1177/0002716212451428

Bennett, W. L., and Livingston, S. (2018). The Disinformation Order: Disruptive Communication and the Decline of Democratic Institutions. *Eur. J. Commun.* 33, 122–139. doi: 10.1177/0267323118760317

Bindra, G. S. (2010). Efficacy of Anti-phishing Measures and Strategies-A research Analysis. *World Acad. Sci. Eng. Technol.* 70, 1409–1415.

Bjola, C., and Papadakis, K. (2020). Digital propaganda, counterpublics and the disruption of the public sphere: the Finnish approach to building digital resilience. *Camb. Rev. Int. Aff.* 2020, 1–29. doi: 10.1080/09557571.2019.1704221

Boutyline, A., and Willer, R. (2017). The social structure of political echo chambers: variation in ideological homophily in online networks. *Polit. Psychol.* 38, 551–569. doi: 10.1111/pops.12337

Brashier, N. M., and Marsh, E. J. (2020). "Judging truth," in *Annu. Rev. Psychol.* Vol. 71, ed. S. T. Fiske (Palo Alto, CA: Annual Reviews), 499–515.

Bratich, J. (2020). Civil society must be defended: misinformation, moral panics, and wars of restoration. *Commun. Cult. Crit.* 13, 311–332. doi: 10.1093/ccc/tcz041

Bronstein, M., Pennycook, G., Bear, A., Rand, D., and Cannon, T. (2019). Belief In Fake News Is Associated With Delusionality, Dogmatism, Religious Fundamentalism, And Reduced Analytic Thinking. *J. Appl. Res. Mem. Cog.* 1, 108–117. doi: 10.1016/j.jarmac.2018.09.005

Burke, K. (1969). *A Rhetoric of Motives*. Berkeley: University of California Press.

Burke, K. (1984). *Permanence and Change: An Anatomy of Purpose*. Berkeley: University of California Press.

Cadwalladr, C., and Graham-Harrison, E. (2018). Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach. *Guardian* 17:22.

Chadwick, A., Vaccari, C., and O'Loughlin, B. (2018). Do Tabloids Poison the Well of Social Media? Explaining Democratically Dysfunctional News Sharing. *New Media Soc.* 20:146144481876968.

Chen, J. (2020). *Social Media Demographics to Inform your Brand's Strategy in 2020*. Available online at: https://sproutsocial.com/insights/new-social-media-demographics/.(accessed on January, 15 2020).

Chetty, N., and Alathur, S. (2018). Hate Speech Review in the Context of Online Social Networks. *Agg. Violent Behav.* 40, 108–118. doi: 10.1016/j.avb.2018.05.003

Clausewitz, C. (1982). *On War*, Vol. 20. London: Penguin UK.

Cockcroft, R., and Cockcroft, S. (2005). *Persuading People: An Introduction to Rhetoric*. London: Palgrave.

Cohen, S. J., Kruglanski, A., Gelfand, M. J., Webber, D., and Gunaratna, R. (2018). Al-Qaeda's propaganda decoded: A psycholinguistic system for detecting variations in terrorism ideology. *Terror. Politic. Violence* 30, 142–171. doi: 10.1080/09546553.2016.1165214

Colliander, J. (2019). "This is fake news": investigating the role of conformity to other users' views when commenting on and spreading disinformation in social media. *Comput. Hum. Behav.* 97, 202–215. doi: 10.1016/j.chb.2019.03.032

Commin, G., and Filiol, E. (2013). "Unrestricted warfare versus traditional warfare: A comparative study," in *Proceedings of the12th European Conference on Information Warfare and Security,* Jyvaskyla, (Finland: ECIWS), 38–44.

Connelly, M., and Welch, D. (eds) (2003). *The Management of Perception: Propaganda, the Media and Warfare, 1900–2002*. London: I. B. Tauris.

Connolly, W. E. (2002). *Identity/Difference: Democratic Negotiations of Political Paradox*. Minneapolis: University of Minnesota Press.

Conroy, N. K., Rubin, V. L., and Chen, Y. (2015). Automatic deception detection: Methods for finding fake news. *Proc. Assoc. Inform. Sci. Technol.* 52, 1–4. doi: 10.1002/pra2.2015.145052010082

Converse, P. E. (1962). Information Flow and the Stability of Partisan Attitudes. *Public Opin. Quart.* 26, 578–599. doi: 10.1086/267129

Crilli, C. (2014). How female anchors really feel about the unspoken dress code for women in TV news, *Bustle*. Available online at: https://www.bustle.com/p/how-female-anchors-really-feel-about-the-unspoken-dress-code-for-women-in-tv-news-10115010 (accessed on Aug. 21, 2018).

Cybenko, G., Giani, A., and Thompson, P. (2002). Cognitive hacking: a battle for the mind. *Computer* 35, 50–56. doi: 10.1109/MC.2002.1023788

Dechene, A., Stahl, C., Hansen, J., and Wanke, M. (2010). The truth about the truth: a meta-analytic review of the truth effect. *Pers. Soc. Psychol. Rev.* 14, 238–257. doi: 10.1177/1088868309352251

De Faveri, C., Moreira, A., and Souza, E. (2017). "Deception planning models for Cyber Security," in *Proceeding of the 17th International Conference on Computational Science and Its Applications (ICCSA)*, (Piscataway,NJ: IEEE), 1-8.

De Keersmaecker, J., and Roets, A. (2017). 'Fake news': incorrect, but hard to correct. The role of cognitive ability on the impact of false information on social impressions. *Intelligence* 65, 107–110. doi: 10.1016/j.intell.2017.10.005

Delli, C., Michael, X., and Keeter, S. (1997). *What Americans Know About Politics And Why It Matters*. New Haven: Yale University Press.

Dey, A. (2018). Fake News Pattern Recognition using Linguistic Analysis. *Elect. Vision* 2018, 305–309.

Digital Shadows. (2018). *The Business of Disinformation: A Taxonomy.* Available online at: https://resources.digitalshadows.com/whitepapers-and-reports/the-business-of-disinformation-fake-news (accessed November 21, 2020).

Dijksterhuis, A. (2004). I like myself but I don't know why: enhancing implicit self-esteem by subliminal evaluative conditioning. *J. Pers. Soc. Psychol.* 86, 345–355. doi: 10.1037/0022-3514.86.2.345

Dorrell, O. (2017). *Brietbart and other alt-right websites darlings of Russian propaganda effort, USA Today*. Available online at: https: //www.usatoday.com/story/news/world/2017/08/24/breitbart-other-alt-right-websites-darlings-russian-propaganda-effort/598258001/ (accessd on Aug, 24 2017).

Downs, A. (1957). *An Economic Theory of Democracy*. 1st ed. Manhattan, NY: Harper and Row.

Easton, D. (1975). A re-assessment of the concept of political support. *Br. J. Polit. Sci.* 5, 435–457. doi: 10.1017/S0007123400008309

Edgerly, S. (2017). Making sense and drawing lines: Young adults and the mixing of news and entertainment. *J. Stud.* 18, 1052–1069. doi: 10.1080/1461670x.2015.1100522

Ellul, J. (1965). *Propaganda: The Formation of Men's Attitudes*. New York, NY: Alfred A. Knopf.

Erfurth, W. (1943). "Surprise," in *The Roots of Strategy, Book 3. Military Classics*, eds S. Possony and D. Vilfroy (Harrisburg, PA: Stagpole Books), 385–547.

Farhall, K., Carson, A., Wright, S., Gibbons, A., and Lukamto, W. (2019). "Political Elites' Use of Fake News Discourse Across Communications Platforms. *Int. J. Commun.* 13:23.

Fenn, E., Ramsay, N., Kantner, J., Pezdek, K., and Abed, E. (2019). Nonprobative photos increase truth, like, and share judgments in a simulated social media environment. *J. Appl. Res. Mem. Cogn.* 8, 131–138. doi: 10.1016/j.jarmac.2019.04.005

Ferejohn, J. A. (1990). "Information and the Electoral Process," in *Information and Democratic Processes*, eds A. John, Ferejohn, H. James, and Kuklinski (Urbana: University of Illinois Press), 1–22. doi: 10.1355/9789814786942-003

Flanagin, A., and Metzger, M. (2014). Digital media and perceptions of source credibility in political communication. *Oxford Handbooks Online* 417. doi: 10.1093/oxfordhb/9780199793471.013.65

Foucault, M. (2003). *Society Must Be Defended": Lectures at the Collège de France, 1975-1976*. Equitable Building, NY: Macmillan.

Garrett, R. K., Nisbet, E. C., and Lynch, E. K. (2013). Undermining the Corrective Effects of Media-Based Political Fact Checking? The Role of Contextual Cues and Naïve Theory. *J. Commun.* 63, 617–637. doi: 10.1111/jcom.12038

Gartzke, E., and Lindsay, J. R. (eds) (2019). *Cross-domain Deterrence: Strategy in an Era of Complexity*. Oxford: Oxford University Press.

Gazzaniga, M. S. (2014). *Handbook of Cognitive Neuroscience*. New York, NY: Springer.

Gerbner, G. (1956). Toward a general model of communication. *Educ. Technol. Res. Dev.* 4, 171–199. doi: 10.1007/BF02717110

Giddens, A. (1991). *Modernity and Self-Identity: Self and Society in the Late Modern Age*. California: Stanford University Press.

Glantz, D. M. (2006). *Soviet Military Deception in the Second World War*. Soviet Military Theory and Practice Series. London: Frank Cass and Co Ltd.

Glazer, A. (1995). "Political Equilibrium under Group Identification," in *Information, Participation, and Choice: An Economic Theory of Democracy in Perspective*, ed. B. Grofman (Ann Arbor: University of Michigan Press), 81–92.

Graves, L. (2017). Anatomy of a Fact Check: Objective Practice and the Contested Epistemology of Fact Checking. *Commun. Cult. Critiq.* 10, 518–537. doi: 10.1111/cccr.12163

Graves, L. (2018). Boundaries Not Drawn. *J. Stud.* 19, 613–631. doi: 10.1080/1461670X.2016.1196602

Grofman, B. (1995). *Information, Participation, and Choice: An Economic Theory of Democracy in Perspective*. Ann Arbor: University of Michigan Press.

Gross, L. (2009). A Broken Trust: Lessons from the Vaccine–Autism Wars: Researchers long ago rejected the theory that vaccines cause autism, yet many parents don't believe them. Can scientists bridge the gap between evidence and doubt? *PLoS Biology* 7:e1000114. doi: 10.1371/journal.pbio.1000114

Gu, L., Kropotov, V., and Yarochkin, F. (2017). The fake news machine: How propagandists abuse the internet and manipulate the public. *Trend Micro* 5, 1–85.

Guess, A., Nagler, J., and Tucker, J. (2019). Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Sci. Adv.* 5:eaau4586. doi: 10.1126/sciadv.aau4586

Handel, M. I. (1987). "Introduction: Strategic and Operational Deception in Historical Perspective," in *Strategic and Operational Deception in the Second World War*, ed. M. I. Handel (London: Frank Cass and Co Ltd).

Hardaker, C., and McGlashan, M. (2016). "Real men don't hate women": Twitter rape threats and group identity. *J. Pragmatics*. 91, 80–93. doi: 10.1016/j.pragma.2015.11.005

Hillman, B. L. (2013). The Clothes I Wear Help Me to Know My Own Power": The Politics of Gender Presentation in the Era of Women's Liberation. *Front. J. Women Stud.* 34:155. doi: 10.5250/fronjwomestud.34.2.0155

Hillyard, S. A., and Münte, T. F. (1984). Selective attention to color and location: An analysis with event-related brain potentials. *Percep. Psychophys.* 36, 185–198. doi: 10.3758/bf03202679

Hochschild, J. L., and Katherine, L. E. (2015). *Do Facts Matter?: Information and Misinformation in American Politics*, 1 Edn. Norman: University of Oklahoma Press.

Hofstede, G. H., Hofstede, G. J., and Minkov, M. (2010). *Cultures and Organizations : Software of the Mind*, 3rd Edn. New York, NY: McGraw-Hill.

Holone, H. (2016). The filter bubble and its effect on online personal health information. *Croat. Med. J.* 57, 298–301. doi: 10.3325/cmj.2016.57.298

Horne, B., and Adali, S. (2017). *This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news*. California: Association for the Advancement of Artificial Intelligence.

Hostetter, A., and Hostetter, E. (2011). Robert Edmond Jones: Theatre and Motion Pictures, Bridging Reality and Dreams. *Theatre Symp.* 19, 26–40. doi: 10.1353/tsy.2011.0003

Information Commissioner's Office. (2018a). *Democracy disrupted? Personal information and political influence*. London: Information Commissioner's Office.

Information Commissioner's Office. (2018b). *Investigation into the use of data analytics in political campaigns: A report to Parliament*. London: Information Commissioner's Office.

Innis, H. A. (2007). *Empire and Communications*. Toronto: Dundurn Press Ltd.

Jamieson, K. H. (2018). *Cyberwar: How Russian Hackers and Trolls Helped Elect a President What We Don't, Can't, and Do Know*. Oxford: Oxford University Press.

Jankowski, N. W. (2018). Researching Fake News: A Selective Examination of Empirical Studies. *Javnost*. *Public* 25, 248–255. doi: 10.1080/13183222.2018.1418964

Jensen, M. (2018). Russian Trolls and Fake News: Information or Identity Logics? *J. Int. Aff.* 71, 115–124. doi: 10.2307/j.ctv3dnq9f.14

Jensen, M., and Chen, T. (forthcoming). Illiberal media in liberal democracy: examining identity in Australia's Mandarin language news. *Issues and Studies*.

Jensen, M., and Sample, C. (2019). *Weaponized narratives and foreign and domestic sources of political influence*. Belgium:: European Consortium for Political Research.

Jones, R. E. (2004). *The Dramatic Imagination: Reflections and Speculations on the Art of the Theatre, Reissue*. Abingdon: Routledge.

Jowett, G., and O'Donnell, V. (1986). *Propaganda and persuasion*. People and communication 18. Newbury Park, CA: SAGE.

Kahneman, D. (2011). *Thinking, Fast and Slow*. New York, NY: Macmillan.

Kahneman, D., and Tversky, A. (1973). On the psychology of prediction. *Psychol. Rev.* 80, 237–251. doi: 10.1037/h0034747

Kalla, J. L., and Broockman, D. E. (2018). The Minimal Persuasive Effects of Campaign Contact in General Elections: Evidence from 49 Field Experiments. *Am. Politic. Sci. Rev.* 112, 148–166. doi: 10.1017/s0003055417000363

Keane, J. (2018). Post-Truth Politics and Why the Antidote Isn't Simply 'fact-Checking' and Truth. Available online at: http://theconversation.com/post-truth-politics-and-why-the-antidote-isnt-simply-fact-checking-and-truth-87364 (accessed on June 25, 2018).

King, G., Pan, J., and Roberts, M. E. (2017). How the Chinese government fabricates social media posts for strategic distraction, not engaged argument. *Am. Politic. Sci. Rev.* 111, 484–501. doi: 10.1017/S0003055417000144

Kuniecki, M., Pilarczyk, J., and Wichary, S. (2015). The color red attracts attention in an emotional context. An ERP study. *Front. Hum. Neurosci.* 9:212. doi: 10.3389/fnhum.2015.00212

Lash, S. (2002). *Critique of Information*. Thousand Oaks: Sage Publications.

Lasswell, H. D. (1927). *Propaganda Technique in the World War*. Cambridge: Ravenio Books.

Lavoipierre, A. (2017). *"Fake News' Named 2016 Word of the Year" ABC News*. Available online at: http://www.abc.net.au/news/2017-01-25/fake-news-named-2016-word-of-the-year/8211056 (February 16, 2018)

Lazer, D. M. J., (2018). The Science of Fake News. *Science* 359, 1094–1096.

Luhmann, N. (1995). *Social Systems*. Stanford: Stanford University Press.

Lynn, W. J. (2010). Defending a New Domain: The Pentagon's Cyberstrategy. *Foreign Aff.* 89, 97–108.

Marcellino, W., Smith, M. L., Paul, C., and Skrabala, L. (2017). *Monitoring Social Media*. Santa Monica, CA: Rand Corporation.

Marquardt, D. (2019). Linguistic Indicators in the Identification of Fake News. *Med. Stud.* 3, 95–114. doi: 10.17951/ms.2019.3.95-114

Mason, L. (2018). *Uncivil Agreement: How Politics Became Our Identity*. Chicago: University of Chicago Press.

Mathiesen, K. (2018). "Fighting Fake News: The Limits of Critical Thinking and Free Speech," in *Information Literacy and Libraries in the Age of Fake News*, ed. E. Denise (Santa Barbara: ABC-CLIO), 77–93.

McAlaney, J., and Benson, V. (2020). *Cybersecurity as a social phenomenon*. Amsterdam: Elsevier, 1–8.

McCumber, J. (1991). "Information systems security: A comprehensive model," in *Proceedings of the 14th National Computer Security Conference*, (Washington, DC: CSRC), 328–337.

McGann, A. J. (2006). *The Logic of Democracy: Reconciling Equality, Deliberation, and Minority Protection*. Michigan: University of Michigan Press.

McKernon, E. (1925). *Fake News and the Public in Harper's Magazine*. Availoable online at: https://harpers.org/archive/1925/10/fake-news-and-the-public/ (accessed on June, 30 2017).

Mena, P. (2019). Principles and Boundaries of Fact-checking: Journalists' Perceptions. *J. Pract.* 13, 657–672. doi: 10.1080/17512786.2018.1547655

Menz, J. (2009). Obama. In his own words: The candidate and the president. Available online at: http://www.diva-portal.org/smash/get/diva2:226983/FULLTEXT01.pdf

Mesoudi, A., Whiten, A., and Dunbar, R. (2006). A bias for social information in human cultural transmission. *Br. J. Psychol.* 97, 405–423. doi: 10.1348/000712605x85871

Meza, S. (2017). *Fake news' named word of the year in Newsweek*. Available online at :http://www.newsweek.com/fake-news-word-year-collins-dictionary-699740 (accessed on November, 2 2017).

Mhamdi, C. (2016). Transgressing Media Boundaries: News Creation and Dissemination in a Globalized World. *Med. J. Soc. Sci.* 7:9462.

Minkov, M. (2013). *Cross-Cultural Analysis*. Thousand Oaks: Sage Publications.

Miskimmon, A., O'Loughlin, B., and Roselle, L. (2014). *Strategic Narratives: Communication Power and the New World Order*. NewYork, NY: Routledge.

Moeslein, A. (2019). *Bombshell costume designer Collen Atwood on what makes a signature fox news look, Glamour*. Available online at: https://www.glamour.com/story/bombshell-costume-designer (accessed on December, 23 2019).

Molina, M. D., Sundar, S. S., Le, T., and Lee, D. (2019). Fake News Is Not Simply False Information: A Concept Explication and Taxonomy of Online Content. *Am. Behav. Sci.* 2019:000276421987822. doi: 10.1177/0002764219878224

Montgomery, M. (2017). Post-truth politics: Authenticity, populism and the electoral discourses of Donald Trump. *J. Lang. Politics* 16, 619–639. doi: 10.1075/jlp.17023.mon

Morgan, E., and Thompson, M. (2018). *Information Warfare: An Emergent Australian Defence Force Capability*. Washington, DC. Discussion Paper. Available online at: https://csis-prod.s3.amazonaws.com/s3fs-public/publication/181023_InformationWarfare.pdf?HDgpzxVHrQayYNZb7t5sJQPlRw9MJCsw (accessd on Oct, 24, 2017).

Mueller, R. S. (2019). *The Mueller Report: Report on the Investigation Into Russian Interference in the 2016 Presidential Election*. Sweden: WSBLD.

Nagpal, S., Singh, M., Singh, R., Vatsa, M., and Nalini, R. (2019). Deep Learning for Face Recognition: Pride or Prejudiced? *arXiv* 2019:1904.01219.

Neuman, S. (2020). *Man dies, woman hospitalized after taking form of chloroquine to prevent COVID-19 in NPR*. Available online at: https://npr.org/sections/coronavirus-live-updates/2020/03/24/820512107/man-dies-woman-hospitalized-after-taking-form-of-chloroquine-to-prevent-covid-19 (accessed on March, 24 2020).

Neilsen, R. K., and Graves, L. (2017). *"News you don't believe": Audience perspectives on fake news* (Reuters Institute for the Study of Journalism Factsheets). Reuters Institute for the Study of Jounalism. Available online at: https://ora.ox.ac.uk/objects/uuid:6eff4d14-bc72-404d-b78a-4c2573459ab8.

Newman, E. J., Garry, M., Bernstein, D. M., Kantner, J., and Lindsay, D. S. (2012). Nonprobative photographs (or words) inflate truthiness. *Psychon. Bull. Rev.* 19, 969–974. doi: 10.3758/s13423-012-0292-0

Nisbett, R. (2010). *The Geography of Thought: How Asians and Westerners Think Differently and Why*. New York: Simon and Schuster.

Northcraft, G. B., and Neale, M. A. (1987). Experts, amateurs, and real-estate - an anchoring-and-adjustment perspective on property pricing decisions. *Organ. Behav. Hum. Decis. Process.* 39, 84–97. doi: 10.1016/0749-5978(87)90046-x

Nyhan, B., and Reifler, J. (2010). When Corrections Fail: The Persistence of Political Misperceptions. *Politic. Behav.* 32, 303–330. doi: 10.1007/s11109-010-9112-2

Nyhan, B., and Reifler, J. (2015). The Effect of Fact-Checking on Elites: A Field Experiment on U.S. State Legislators. *Am. J. Politic. Sci.* 59, 628–640. doi: 10.1111/ajps.12162

Oswald, M. E., and Grosjean, S. (2004). *Confirmation bias. Cognitive illusions: A handbook on fallacies and biases in thinking, judgement and memory*. London: Psychology Press.

Pathak, A., and Srihari, R. K. (2019). "BREAKING! Presenting Fake News Corpus For Automated FactChecking," in *Proceedings of the 57th Annual*

Meeting of the Association for Computational Linguistics: Student Research, (Stroudsburg: Workshop.Association for Computational Linguistics), 357–362.

Payne, A. (2017). *Russia used a network of 150,000 Twitter accounts to meddle in Brexit.in Business Insider*. Available online at: https://amp.businessinsider.com/russia-used-twitter-accounts-to-meddle-in-brexit-investigation-shows-2017-11?__twitter_impression=true (accessed on Nov, 15 2017).

Pennycook, G., and Rand, D. (2019). Lazy, Not Biased; Susceptibility To Partisan Fake News Is Better Explained By Lack Of Reasoning Than By Motivated Reasoning. *Cognition* 188, 39–50. doi: 10.1016/j.cognition.2018.06.011

Pérez-Rosas, V., Kleinberg, B., Lefevre, A., and Mihalcea, R. (2018). "Automatic Detection of Fake News," in *Proceedings of the 27th International Conference on Computational Linguistics*, (Santa Fe: Association for Computational Linguistics), 3391–3401.

Phaedrus (2008). *The Fables of Paedrus. Fable IV. Prometheus and Cunning in Translated by Riley and Smart, Project Gutenberg*. Available online at: https://www.gutenberg.org/files/25512/25512-h/25512-h.htm#NF_IV (accessed on May, 18 2008). 438.

Plotkina, D., Munzel, A., and Pallud, J. (2020). Ilusions of truth – experimental insights into human and algorithmic detections of fake online reviews. *J. Bus. Res.* 109, 511–523. doi: 10.1016/j.jbusres.2018.12.009

Pomerantsev, P., and Weiss, M. (2014). The menace of unreality: How the Kremlin weaponizes information, culture and money. *Interpreter* 2014:22.

Pratkanis, A. R., and Aronson, E. (1992). *Age of propaganda: The everyday use and abuse of persuasion*. New York: W. H. Freeman.

Risley, J. (2016). *"Microsoft's millenial chatbot Tay.ai pulled offline after Internet teaches her racism", Geek Wire*. Available online at: https://www.geekwire.com/2016/even-robot-teens-impressionable-microsofts-tay-ai-pulled-internet-teaches-racism/ (accessed on March, 24 2016).

Risso, L. (2018). Harvesting your soul? Cambridge analytica and brexit. *Brexit Means Brexit* 2018, 75–90.

Roozenbeek, J., and Van Der Linden, S. (2019). The fake news game: actively inoculating against the risk of misinformation. *J. Risk Res.* 22, 570–580. doi: 10.1080/13669877.2018.1443491

Rosenblatt, S. (2018). *Exacerbating our fake news problem: Chatbots*. Available online at: https://www.the-parallax.com/2018/03/26/fake-news-chatbots/. (accessed on March, 26 2018).

Rushkoff, D. (2013). *Present Shock: When Everything Happens Now*. New York, NY: Penguin.

Sample, C. (2015). *Cyber + Culture Early Warning Study*. Pittsburgh, PA: Carnegie Mellon University.

Sample, C., Justice, C., and Darraj, E. (2018). "A Model for Evaluating Fake News," in *Proceedings from NATO CyConUS Conference*, (Washington, DC: NATO).

Schafer, B. (2018). *A View from the Digital Trenches - Lessons from Year One of Hamilton 68*. Washington, DC: German Marshall Fund.

Schapals, A. K. (2018). Fake News. *J. Pract.* 12, 976–985. doi: 10.1080/17512786.2018.1511822

Schwarz, N. (2012). "Feelings-as-information theory," in *Handbook of Theories of Social Psychology*, Vol. 1, eds P. A. M. Van Lange, A. W. Kruglanski, and E. T. Higgins (Thousand Oaks, CA: Sage Publications Ltd), 289–308. doi: 10.4135/9781446249215.n15

Scott, I. (2006). From Toscanini to Tennessee: Robert Riskin, the OWI and the Construction of American Propaganda in World War II. *J. Am. Stud.* 40, 347–366. doi: 10.1017/s0021875806001411

Shakespeare, W. (2009). *As You Like It*. Cambridge: Cambridge University Press.

Shannon, C. (1948). "A Mathematical Theory of Communication.". *Bell Syst. Tech. J.* 27:125.

Sheppard, S. I. (2007). *The Partisan Press: A History of Media Bias in the United States*. US: McFarland.

Shu, K., Wang, S., Tang, J., and Liu, H. (2017). Fake news detection on social media: A data mining perspective. *ACM SIGKDD. Explor. Newslett.* 19, 22–36. doi: 10.1145/3137597.3137600

Shukrun-Nagar, P. (2009). Quotation markers as intertextual codes in electoral propaganda. *Text Talk Int. J. Lang. Dis. Commun. Stud.* 29, 459–480. doi: 10.1515/text.2009.024

Sides, J., Tesler, M., and Vavreck, L. (2018). *Identity Crisis: The 2016 Presidential Campaign and the Battle for the Meaning of America*. Princeton, NJ: Princeton University Press.

Singer, P. W., and Brooking, E. T. (2018). *LikeWar: The Weaponization of Social Media*. New York, NY: Houghton Mifflin.

Sîrbu, A., Pedreschi, D., Giannotti, F., and Kertész, J. (2019). Algorithmic bias amplifies opinion fragmentation and polarization: a bounded confidence model. *PLoS One* 14:e0213246. doi: 10.1371/journal.pone.0213246

Snopes. (2020). *Poisoned Halloween Candy*. Available online at: https://www.snopes.com/fact-check/halloween-non-poisonings/ (accessed November 21, 2020).

Springer, M. (2020). *Social media firms fail to act on Covid-19 fake news in BBC*. Available online at: https://www.bbc.com/news/technology-52903680 (accessed on June, 4 2020).

Stanislavski, C. (1989). *An Actor Prepares*. NewYork, NY: Routledge.

Stubbersfield, J. M., Tehrani, J. J., and Flynn, E. G. (2015). Serial killers, spiders and cybersex: Social and survival information bias in the transmission of urban legends. *Br. J. Psychol.* 106, 288–307. doi: 10.1111/bjop.12073

Szfranski, R. (1997). *A theory of information warfare: Preparing for 2020*. Maxwell: Air University Maxwell Airforce Base.

Tandoc, E. C., Zheng, W. L., and Ling, R. (2018). Defining 'Fake News. *Digit. Journal.* 6, 137–153.

Taylor, C. (1992). *Sources of the Self: The Making of the Modern Identity*. Cambridge: Cambridge University Press.

Tenove, C., Buffie, J., McKay, S., and Moscrop, D. (2018). *Digital Threats to Democratic Elections: How Foreign Actors Use Digital Techniques to Undermine Democracy*, Research Report. Columbia: University of British Columbia.

Tocqueville, A. (2010). . *Democracy in America: In Four Volumes*. Bilingual. Indianapolis: Liberty Fund Inc.

Travers, M., Van Boven, L., and Judd, C. (2014). The secrecy heuristic: inferring quality from secrecy in foreign policy contexts. *Polit. Psychol.* 35, 97–111. doi: 10.1111/pops.12042

Tucker, J., Buffie, J., McKay, S., and Moscrop, D. (2018). *Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature*. Rochester. New York, NY: Social Science Research Network.

Tversky, A., and Kahneman, D. (1983). Judgment under uncertainty: heuristics and biases. *Soc. Behav. Sci.* 14:22.

Unkelbach, C., Koch, A., Silva, R. R., and Garcia-Marques, T. (2019). Truth by repetition: Explanations and implications. *Curr. Direc. Psychol. Sci.* 28, 247–253. doi: 10.1177/0963721419827854

Uscinski, J. E. (2015). The Epistemology of Fact Checking (Is Still Naìve): Rejoinder to Amazeen. *Critic. Rev.* 27, 243–252. doi: 10.1080/08913811.2015.1055892

Van der Linden, S., Leiserowitz, A., Rosenthal, S., and Maibach, E. (2017). Inoculating the public against misinformation about climate change. *Glob. Challenges* 1:1600008. doi: 10.1002/gch2.201600008

Verrall, N., and Mason, D. (2019). The Taming of the Shrewd. *RUSI J.* 163, 20–28. doi: 10.1080/03071847.2018.1445169

Virilio, P., and Sylvère, L. (2008). *Pure War*. Los Angeles: Semiotexte.

Volodymyr, H. (2016). The Hybrid Warefare. *Ontology* 2011:26.

Vosoughi, S., Roy, D., and Aral, S. (2018). The spread of true and false news online. *Science* 359, 1146–1151. doi: 10.1126/science.aap9559

Vraga, E. K., and Bode, L. (2017). I Do Not Believe You: How Providing a Source Corrects Health Misperceptions across Social Media Platforms. *Inform. Commun. Soc.* 1, 1–17. doi: 10.1080/10410236.2020.1775447

Waisbord, S. (2018). Truth is What Happens to News. *J. Stud.* 19, 1866–1878. doi: 10.1080/1461670X.2018.1492881

Wakefield, A. J. (1999). MMR vaccination and autism. *Lancet* 354, 949–950. doi: 10.1016/s0140-6736(05)75696-8

Waltz, E. (2008). "Know Thy Enemy: Acquisition, Representation, and Management of Knowledge About Adversary Organizations," in *Information Warfare and Organizational Decision-Making*, ed. A. Kott (Norwood: Artech House).

Whaley, B. (2007). *Strategem–Deception and Surprise in War*. London: Artech House.

Wittgenstein, L. (1967). *Zettel*. Berkeley: University of California Press.

Younger, A. (2018). *MI6 'C' speech on fourth generation espionage*. Available on line at: https://www.gov.uk/government/speeches/mi6-c-speech-on-fourth-generation-espionage (accessed on December, 03 2018).

Zaharna, R. S. (2018). *New Realities in Foreign Affairs: Diplomacy in the 21st Century*. Berlin: German Institute for International and Security Affairs.

Zebrowitz, L. A., and Franklin, R. G. Jr. (2014). The attractiveness halo effect and the babyface stereotype in older and younger adults: similarities, own-age accentuation, and older adult positivity effects. *Exp. Aging Res.* 40, 375–393. doi: 10.1080/0361073X.2014.897151

Zhou, X., and Zafarani, R. (2018). Fake News: A Survey of Research, Detection Methods, and Opportunities. *ACM Comput. Surv.* 1, 1–40. doi: 10.1145/3395046

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Two researchers were funded by the US Army Research Laboratory, as listed in section "Author Contributions."

frontiers in Psychology

REVIEW
published: 18 June 2021
doi: 10.3389/fpsyg.2021.561011

The Role of User Behaviour in Improving Cyber Security Management

Ahmed A. Moustafa[1,2,3]*, *Abubakar Bello*[4] *and Alana Maurushat*[4]

[1] School of Psychology, Western Sydney University, Sydney, NSW, Australia, [2] The Marcs Institute for Brain, Behaviour and Development, Western Sydney University, Sydney, NSW, Australia, [3] Department of Human Anatomy and Physiology, Faculty of Health Sciences, University of Johannesburg, Johannesburg, South Africa, [4] School of Social Sciences, Western Sydney University, Sydney, NSW, Australia

Information security has for long time been a field of study in computer science, software engineering, and information communications technology. The term 'information security' has recently been replaced with the more generic term cybersecurity. The goal of this paper is to show that, in addition to computer science studies, behavioural sciences focused on user behaviour can provide key techniques to help increase cyber security and mitigate the impact of attackers' social engineering and cognitive hacking methods (i.e., spreading false information). Accordingly, in this paper, we identify current research on psychological traits and individual differences among computer system users that explain vulnerabilities to cyber security attacks and crimes. Our review shows that computer system users possess different cognitive capabilities which determine their ability to counter information security threats. We identify gaps in the existing research and provide possible psychological methods to help computer system users comply with security policies and thus increase network and information security.

Keywords: cyber security, social engineering, information security, phishing, cognitive hacking

OPEN ACCESS

Edited by:
Richard Keith Wortley,
University College London,
United Kingdom

Reviewed by:
Karima Boudaoud,
Université Côte d'Azur, France
Palvi Aggarwal,
Carnegie Mellon University,
United States

****Correspondence:***
Ahmed A. Moustafa
a.moustafa@westernsydney.edu.au

Specialty section:
This article was submitted to Cognition, a section of the journal Frontiers in Psychology

Received: *11 May 2020*
Accepted: *03 May 2021*
Published: *18 June 2021*

Citation:
Moustafa AA, Bello A and Maurushat A (2021) The Role of User Behaviour in Improving Cyber Security Management. Front. Psychol. 12:561011. doi: 10.3389/fpsyg.2021.561011

INTRODUCTION

According to National Initiative for Cybersecurity Careers and Studies, cybersecurity is defined as 'the activity or process, ability, or capability or state whereby information and communications systems and the information contained therein are protected from and/or defended against damage, unauthorised use or modification, or exploitation.' Cyber and network systems involve at least four components: computer system users, security system analysts, cyber attackers, and computer systems. Cyber attackers often attempt to obtain, modify, or keep unauthorised information (Landwehr, 1981; Thompson, 2004).

Most of the research on cybersecurity has focused on improving computer network systems (Nobles, 2018), as many believe that information technology advances and software development is the main way to increase information security (Sadkhan, 2019; Benson and Mcalaney, 2020). Fewer studies have been conducted on enhancing cognitive capabilities and situational awareness of system analysts (D'Amico et al., 2005; Barford, 2010; Dutt et al., 2013; Knott et al., 2013; Tyworth et al., 2013; Mancuso et al., 2014; Gutzwiller et al., 2015; Aggarwal et al., 2018; Veksler et al., 2018).

However, cyber attackers can also manipulate the minds of computer system users, rather than a computer system itself, by, for example, using social engineering (e.g., tricking of computer system users to gain information, such as passwords) and cognitive hacking (e.g., spreading of misinformation) to break into a network or computer system (Cybenko et al., 2002; Thompson, 2004; McAlaney et al., 2015; King et al., 2018; Fraunholz et al., 2019). According to

Bowen et al. (2014), social engineering attacks account for 28% of total cyber security attacks and 24% of these attacks occurred due to phishing. According to CyberEdge Reports, more than 70% of social engineering attacks have been successful in the last few years. In the 2018 and 2019 reports by Telstra, human errors are the greatest threat in cybersecurity. The reports claim that phishing (and spear-phishing) attacks were the most common attacks and they utilised partial social engineering and fraud to scam victims into installing malware or illegitimate websites to acquire their credentials. In these types of attacks, victims are often sent emails or text messages that appear, for example, to be for a software upgrade, legitimate correspondence from a third party supplier, information on a current storm or crisis, or notifications from a bank or a social networking site. In addition to falling victim to phishing attacks, computer system users also conduct other cyber security errors, such as sharing passwords with friends and family and also not installing software updates.

It is important to note that there are individual differences among computer system users in terms of complying with security behaviours. Several studies found that individual differences in procrastination, impulsivity, future thinking, and risk taking behaviours can explain differences in complying with security policies. Importantly, given the existing human errors that can impact network security, we will discuss the use of psychological methods to improve compliance with security policies. Such psychological methods include using novel polymorphic security warnings, rewarding and penalizing good and bad cyber behaviour, and increasing thinking about future consequence of actions.

This paper is structured as follows. First, we discuss studies and measures related to complying with security policies. Second, we discuss kinds of cyber security errors done by many computer system users, including falling victim to phishing, sharing passwords, and not installing software updates and. Third, we discuss individual differences underlying cyber security behaviours in computer system users, including procrastination, impulsivity, future thinking, and risk taking behaviours. We conclude by suggesting psychological methods that could be used to move user behaviour toward secure practices.

COMPLYING WITH SECURITY POLICIES

Complying with security policies is one key behaviour to protect computer and network systems. There have been few studies on the psychology of compliance with security policies (Chan et al., 2005; Lee and Kozar, 2005; Hazari et al., 2009; Anderson and Agarwal, 2010; Maurushat, 2010; Guo et al., 2011). A lack of complying with security policies can significantly undermine information security (Greenwald et al., 2004; Mishra and Dhillon, 2006; West, 2008). For example, several studies have shown that computer system users often ignore security warnings (Schechter et al., 2007; Akhawe and Felt, 2013; Bravo-Lillo et al., 2013; Brase et al., 2017).

To measure such humans' security behaviours, Egelman and Peer (2015) developed the Security Behaviour Intentions scale. The scale measures attitudes toward choosing passwords, device security, regularly updating software, and general awareness about security attacks. The scale has 16 questions, such as (a) I use a password/passcode to unlock my laptop or tablet, (b) When I'm prompted about a software update, I install it right away, (c) I manually lock my computer screen when I step away from it, and (d) If I discover a security problem, I continue what I was doing because I assume someone else will fix it. The scale itself represents very basic aspects of security protection and mitigation techniques. As we discuss below, several studies have used this scale to measure types of security errors done by computer system users.

Non-compliance with a security policy can go beyond mere ignoring warnings, choosing poor passwords or failing to adopt recommended security measures. In a recent study, Maasberg et al. (2020) found that the dark triad traits (machiavellianism, narcissism and psychopathy, machiavellianism, narcissism and psychopathy, Paulhus and Williams, 2002) correlate with malicious behaviour intentions such as insider threats. Harrison et al. (2018) recently reported that the Dark triad can explain unethical behaviour such as committing cyber fraud. The concept of Dark Triad and Big Five Methods will be explored and critiqued further in the following section.

HUMAN CYBER SECURITY ERRORS

In this section, we describe the kinds of cyber security errors conducted by many computer system users. Several reports have shown that humans are considered the greatest vulnerability to security (Schneier, 2004; Furnell and Clarke, 2012), which has been also confirmed by recent reports. One report estimated that 95% of cyber and network attacks are due to human errors (Nobles, 2018). In our context, humans are either computer system users or security analysts (King et al., 2018; Andrade and Yoo, 2019), though most research on this area focuses on errors done by computer system users. According to Ifinedo (2014), company employees are the weakest link in ensuring system security (for discussion and analysis, also see Sasse et al., 2004; Vroom and von Solms, 2004; Stanton et al., 2005; Guo et al., 2011).

Some human errors related to cyber and network security include, but not limited to, sharing passwords, oversharing information on social media, accessing suspicious websites, using unauthorised external media, indiscriminate clicking on links, reusing the same passwords in multiple places, opening an attachment from an untrusted source, sending sensitive information via mobile networks, not physically securing personal electronic devices, and not updating software (Boyce et al., 2011; Calic et al., 2016). Along these lines, one main issue underlying information and cyber security is the dilemma of increasing availability and ease to access a network or data but, at the same time, maintain security (Veksler et al., 2018). To increase security, organisations often require computer system users to have complex passwords, which makes usability quite difficult. Computer system users, however, tend to take the path of least resistance, such as using a weak password and using the same password for several websites. Below, we discuss prior studies on three kinds of human security errors: falling victim to phishing, sharing passwords with others, and installing software updates.

Falling victim to phishing: Some phishing studies have used a laboratory-based phishing experiment (Jakobsson and Ratkiewicz, 2006; Jagatic et al., 2007). The use of laboratory-based phishing experiment has been shown in a recent study to relate to real-life phishing (Hakim et al., 2020). One study found that over 30% of government employees click on a suspicious link in this phishing email, and many of these have provided their passwords (Baillon et al., 2019). In another study using a similar phishing experiment, around 60% of university students clicked on suspicious link in a phishing email (Diaz et al., 2018). Accordingly, several studies suggest that human factors, behavioural studies, and psychological research must be considered in cyber and network security studies (Hamill and Deckro, 2005; Jones and Colwill, 2008). In another study, Bowen et al. (2014) studied how Columbia University students and academic staff respond to phishing emails, and found that it took people around 4 rounds to discover they are receiving phishing emails.

One recent study also found that a successful phishing attack is related to the Dark Triad traits of the computer users, including machiavellianism, narcissism, and psychopathy (Curtis et al., 2018). In this study, it was found that high scores in narcissism is related to a higher tendency to fall victim to phishing attempts. Along these lines, it was found that neuroticism is related to falling victim to phishing attacks (Halevi et al., 2013). In another study by Gonzalez and colleagues (Rajivan and Gonzalez, 2018), it was found that the use of some cyberattack strategies, such as sending excessive amount of notification and expressing shared interest, were more related to successful phishing.

One study found that even warning people about phishing does not change their response to phishing emails (Mohebzada et al., 2012). Using the Human Aspects of Information Security Questionnaire (HAIS-Q) (Calic et al., 2016; Parsons et al., 2017), it was found that individuals who scored high on the HAIS-Q performed better on a laboratory-based phishing experiment, in which a randomly selected sample of participants (from a firm, university, school, or so) are unknowingly sent a phishing email that urges them to share their password. Herath and Rao (2009) found that computer system users generally underestimate the probability of security breaches and cybercrimes happening to them.

Sharing passwords: Sharing passwords with friends and family, and even strangers is a prevalent example of human cyber security errors. According to Whitty et al. (2015), older adults who score high on perseverance and self-monitoring are more likely to share passwords. Sharing passwords may lead to financial exploitation of older adults, which is among the most common forms of abuse (Bailey et al., 2015). This is the case as many older adults are very trusting of others and strangers, especially on the internet. Like older adults, younger adults also share passwords, especially ones for streaming systems. Younger users (who had grown up with computers) perceived security as an obstacle they had to work around (Smith, 2003). Sharing passwords is generally problematic as most people often use the same passwords for several websites, and thus by sharing a password, others can access their other secure information. One problem with using the same password in many systems is that cybercriminals, once find these passwords in one system, can use these passwords in many other websites.

Installing software updates: One common error underlying cybersecurity behaviours is a delay in or even not at all installing software updates (Rajivan et al., 2020). Using an experimental behavioural decision making study, Rajivan et al. (2020) found that risk-taking behaviours can partly explain some individuals behaviours regarding installing software updates, such that individuals who are more risk taking tend to delay the installation of software updates. Unlike sharing passwords and phishing, the area of installing software updates has not received much attention in the field.

INDIVIDUAL DIFFERENCES UNDERLYING CYBER SECURITY BEHAVIOURS

Individual differences in personality, cognitive and behavioural traits are related to cyber security behaviours. Dawson and Thomson (2018) argue that individual differences in cognitive abilities and personality traits can play a key role in success to secure computer and information systems. Below, we discuss some of these psychological traits.

Procrastination: Complying with security policies is possibly related to cognitive processes, such as working hard to achieve certain goals. One scale, known as "the need for cognition" scale measures working hard, enjoying and participating in activities that require efforts and thinking (Lin et al., 2016). Along these lines, Egelman and Peer (2015) found that performance in the Security Behaviour Intentions Scale is related to the Need for Cognition (NFC), which refers to inclination to exerting cognitive efforts (Cacioppo et al., 1984). Interestingly, a new study has developed a scale to measure procrastination in children and adolescents, which is suitable for the increasing number of young internet users (Keller et al., 2019). Along these lines, Shropshire et al. (2006) reported a link between the intent to comply with information security protocols and conscientiousness (i.e., doing work thoroughly and accurately) (McBride et al., 2012). Further, using the General Decision-Making Style (GDMS) scale (Scott and Bruce, 1995), Egelman and Peer (2015) found that performance in the Security Behaviour Intentions Scale is related to procrastination, such that, individuals who procrastinate were less likely to follow security policies. This is plausible as procrastination is negatively correlated with active participation in activities (Sarmany-Schuller, 1999).

Impulsivity: Complying with security policies may be also related to individual differences in impulsive behaviours. Egelman and Peer (2015) found that performance in the Security Behaviour Intentions Scale is related to Barratt Impulsiveness Scale scores (Patton et al., 1995). Another study found that internet addiction and impulsivity predicts risky cyber behaviours (Hadlington, 2017). Along these lines, Hu et al. (2015)

found that individual differences in self and cognitive control (a key feature of impulsive behaviours) is related to violation of information security policies. Wiederhold (2014) also found that people fall victim to cybersecurity attacks in the pursuit of immediate gratification. One key feature related to impulsivity is thinking about future consequences of one's actions (e.g., saving money now to buy a house in the future vs. spending all money now to enjoy life).

Future thinking: Importantly, complying with security policies may also be related to thinking about the future as well as impact of present actions on future consequences (A. A. Moustafa et al., 2018a; Moustafa et al., 2018b). In other words, individuals who think more about the future may abide by security rules to make sure their computer system is safe in the future. Along these lines, Egelman and Peer (2015) found that performance in the Security Behaviour Intentions Scale is related to Consideration for Future Consequences (CFC) (Joireman et al., 2012). This scale includes items that are very relevant to cyber security behaviours, such as 'I consider how things might be in the future, and try to influence those things with my day to day behaviour', 'I think it is important to take warnings about negative outcomes seriously even if the negative outcome will not occur for many years', and 'When I make a decision, and I think about how it might affect me in the future'.

Risk taking behaviours: Another personality trait related to cyber security is risk taking behaviours. Some studies have found that computer system users who are high in risk taking may be more likely to fall victims to cybercrimes (Henshel et al., 2015; King et al., 2018). Risk is defined as engaging in a behaviour with an uncertain outcome, usually for the benefit of gaining more (Saleme et al., 2018). For example, robbing a bank is risky, as one may get caught. A lack of complying with security policies is risky as the benefit is not doing any additional work, such as software update (which is rewarding), but the risk is falling victim to cybercrimes and phishing. Another example is finding out that there has been a data breach where your personal information such as your username and password has been compromised, but then not doing anything to change your password. The dilemma computer system users face is doing additional work to secure their network or computer systems (too much work but more safe) or not (less work but less safe). Importantly, Egelman and Peer (2015) found that performance in the Security Behaviour Intentions Scale is related to performance in the Domain-Specific Risk-Taking Scale, which has items on general risk taking behaviours in everyday life (Blais and Weber, 2006; Saleme et al., 2018; Saleme and Moustafa, 2020). In several studies, by using the Risky Cybersecurity Behaviours Scale, Security Behaviours Intentions Scale (SeBIS), and Attitudes toward cybersecurity and cybercrime in business (ATC-IB), Hadlington and colleagues (Hadlington, 2017; Hadlington and Murphy, 2018) found that heavy media multitasking is associated with risky cybersecurity behaviours and increased cognitive errors.

Optimism bias is related to risk-based decision making. There have few psychology studies on optimism bias in humans (West, 2008; Sharot, 2011; Moutsiana et al., 2013; Garrett and Sharot, 2017). Generally, people assume that the best will happen to them, and they do not think they are at risk (West, 2008), that is, humans tend to be more optimistic and discount the likelihood of negative events happening to them. For example, people generally do not assume they will have cancer disease, and often discount the likelihood of it happening. This is relevant to research on the psychology of cyber and network security as computer system users may tend to discount the impact of cyber-attacks or crimes happening to them. For example, one study found that people fall victim to cybersecurity attacks due to optimism bias (Wiederhold, 2014). Importantly, future work should investigate individual differences in optimism bias and its relationship to risky cybersecurity behaviours.

Other areas of study that have examined individual differences in cybersecurity are considered under the framework of the Dark Triad and the Big Five Model. The majority of these studies are in the field of cyber bullying which falls outside of the scope of this paper, but other studies have been incorporated into sections of this paper (West, 2008; Goodboy and Martin, 2015; Jacobs et al., 2015; Alonso and Romero, 2017; Rodriguez-Enriquez et al., 2019; Curtis et al., 2021). The Big Five Scale has also been used in cybersecurity and psychology studies. The Big Five Scales refers to Agreeableness, Neuroticism, Openness, Conscientious and Extraversion. We have found, however, that the literature refers to only Neuroticism, Openness and Extraversion. Instead of examining the individual differences of the limited approach of the dark triad and the Big Five Scales we have instead pulled out the multi-dimensional aspects involved with the triad. For example, impulsivity is one component that expands across the different indexes of measurement. The other factors are grouped in **Table 1**.

In sum, in this section, we reviewed prior studies showing that personality traits and individual differences in procrastination, impulsivity, and risk-taking behaviours, are related to cyber security behaviours.

IMPROVING SECURITY BEHAVIOURS USING PSYCHOLOGICAL METHODS

As discussed above, cyber attackers often use social engineering and cognitive hacking methods to break into a network or computer systems (Cybenko et al., 2002; Thompson, 2004; McAlaney et al., 2015; King et al., 2018; Fraunholz et al., 2019). Some computer system users may have some personality traits that make them likely to fall victims to phishing. Accordingly, it is important to equip vulnerable computer system users (i.e., those who may not comply with security policies) with capabilities to mitigate these effects. In this section, we discuss several psychological methods to increase compliance with security policies.

Using novel polymorphic security warnings: According to Anderson et al. (2015), most people ignore security warnings on the internet due to habituation. In the field of psychology, habituation refers to a decreased response to repeated exposure to the same stimulus over time (Rankin et al., 2009). That is, we do not pay attention to objects that we repeatedly see. West (2008) also argued that most warning messages are similar to other message dialogs. Accordingly, computer system users

TABLE 1 | Summary of individual traits founds in applicable theories and instruments.

Individual trait	Test/theory	Instrument
Procrastination	Big Five:	Hunter and Schmidt Meta-Analysis Procedure
	Neuroticism	Academic Procrastination Scale
	Dark Triad:	Adult Inventory of Procrastination
	Machiavellianism	Aitken Procrastination Inventory
	and Psychopathy	Decisional Procrastination Questionnaires
		General Procrastination Scale
		Procrastination Assessment Scale—Students
		Procrastination Log—Behaviour
		Procrastination Self-Statement Inventory
		Test Procrastination Questionnaire
Impulsiveness	Dark Triad:	Hadlington's Examination
	Psychopathy	Abbreviated Impulsiveness Scale
	Narcissism	Barratt's Impulsiveness Scale
	Big 5 Scales:	Security Behaviours Intentions Scale (SeBIS)
	Openness	Ecological Momentary Assessment
	Extraversion	Dysfunctional Impulsivity subscale of the Dickman
		Impulsivity Inventory
Future thinking		Internet Addiction Test
		Wishful Thinking Scale
		Automatic Thoughts Questionnaire
		Entrepreneurial Self-Efficacy (ESE) scale
		Cyber Bullying Attitude Scale
		Cybersecurity Attitudes Scale
Risk taking		Security Behaviour Intentions Scale
		Domain Specific Risk Taking Scale
		Risky Cybersecurity Behaviours Scale

often ignore them, as our brain is not likely to show novelty and attentional allocation response to such security warnings (Moustafa et al., 2009).

According to Wogalter (2006), the use of different polymorphic security warnings over time will help increase attention to these warnings. Along these lines, Anderson et al. (2015) found that the use of polymorphic warnings did not lead to habituation, that is, computer system users can still pay attention and respond to these security warnings. Similar findings were also found by Brustoloni and Villamarín-Salomón (2007). Responding to novel and anomalous activities are aspects of situational awareness, and key for detecting phishing attempts in a cyber or network systems (D'Amico et al., 2005; Barford, 2010; Dutt et al., 2013; Knott et al., 2013; Tyworth et al., 2013; Mancuso et al., 2014; Aggarwal et al., 2018; Veksler et al., 2018). Software engineers should develop attention-capturing security warnings and not standard message dialogs, and these also should change over time in order to increase alertness and attention in computer system users. Using unique and novel security messages is important, as research have reported that these messages can increase brain activation and attentional processes (Moustafa et al., 2009, 2010; Kar et al., 2010).

In addition, other studies have compared security warning design differences between Firefox, Google and Internet Explorer browsers (Akhawe and Felt, 2013). Akhawe and Felt found that browser security warnings can be effective security mechanisms although there were a number of important variables that contribute to click through rates after warnings including warning type, number of clicks, warning appearance, certificate pinning and time spent on warnings.

Rewarding and penalizing good and bad cyber behaviour: In everyday life, we learn from negative (e.g., loss, penalties, etc.) or positive (e.g., reward) outcomes. Humans are often motivated to do certain actions to receive reward and avoid negative outcomes (Frank et al., 2007; Moustafa et al., 2008, 2013, 2015, 2017; Bodi et al., 2009; Piray et al., 2014; Myers et al., 2016). However, in the case of cyber security behaviours, the reward is that nothing bad will happen; that is, the user's computer system will not be attacked if they comply with security policies. In other words, complying with cyber security behaviours is an example of negative reinforcement in which actions (i.e., complying with cyber security policies) prevent the occurrence of a negative outcome (Sidman, 2006; May et al., 2020).

Based on these findings, the use of more concrete rewards and losses may increase compliance with security policies. For example, companies should enforce fines (kind of punishment learning) on employees who do not adhere to security policies and reward ones who do. Maqbool et al. (2020) argued that penalizing individuals should increase security behaviours. Along these lines, Baillon et al. (2019) used a phishing experiment (in which participants click on a link which then ask them to provide their passwords) to study how simulated experience with prior phishing can impact future behaviour. They found that experiencing simulated phishing (i.e., a negative outcome) increases compliance with security policies in the computer system users. It has been found that providing information about the prevalence of phishing (i.e., negative outcome can occur to people) can decrease clicking on suspicious links in phishing emails (Baillon et al., 2019). Accordingly, computer system users should be provided with simulated experience of negative outcomes that may occur due to their erroneous cyber security policies. Further, future studies should explore whether rewarding compliance with security policies will increase future pro security behaviours (Regier and Redish, 2015).

Along these lines, according to Tversky and Kahneman (1986), most people prefer a certain small reward over uncertain big reward, but people prefer uncertain loss than a certain loss (for discussion, also see for discussion, also see Herzallah et al., 2013). In other words, people generally prefer to gamble on losses. This is evident in security behaviours. Given that the reward related to security behaviours is not direct (i.e., nothing bad will happen), using a strong reward should increase adherence to security behaviours. Future research should also investigate the relationship between individual differences in response to rewarding and penalizing outcomes and compliance with security behaviours.

Increasing thinking about future consequence of actions: As mentioned above, some of the key features about lack of complying with cyber security policies is not thinking much about future consequences. It has been found that thinking about future consequences is related to reflective decision making and planning (Eskritt et al., 2014) and can decrease impulsive behaviours, which is related to risky behaviours on the web as we discussed above (Bromberg et al., 2015, 2017). Accordingly, using psychological methods to increase thinking about future consequences of actions can help increase reflective decision making, and thus improve cyber security behaviours (Altintas et al., 2020).

CONCLUSION AND FUTURE DIRECTIONS

Our review shows that some personality traits, such as impulsivity, risk taking, and lack of thinking about future consequences of actions, are related to a lack of compliance with cyber and network security policies. Future research should focus on developing a battery of tests to integrate personality traits and cognitive processes related to cyber and network security behaviours in one framework. This battery of tests should include cognitive processes discussed above, including impulsivity, risk taking, and thinking about future consequences of actions. Furthermore, here, we show that some psychological methods can increase pro-security behaviours, such as rewarding and penalizing security-related behaviours, using novel polymorphic security warnings, and using psychological methods to increase thinking about future consequences of actions. In addition, there are cognitive training methods, including working memory training, that help reduce impulsivity, risk taking and procrastination in the general population (Rosenbaum et al., 2017; Peckham and Johnson, 2018). Such cognitive training methods can be used to ameliorate these behavioural traits and help improve cybersecurity behaviours.

As discussed above, there are different kinds of human errors that can undermine computer and security systems, including sharing passwords, oversharing information on social media, accessing suspicious websites, using unauthorised external media, indiscriminate clicking on links, reusing the same passwords in multiple places, using weak passwords, opening an attachment from an untrusted source, sending sensitive information via mobile networks, not physically securing personal electronic devices, and not updating software. However, most of the research conducted on human errors has been on phishing emails and sharing passwords. Future research should also investigate individual differences and contextual information (e.g., mood status, urgency at work, or multitasking) underlying other kinds of cyber security errors, such as using same or weak passwords in several websites, not connecting with virtual private networks and not encrypting data.

There are computational cognitive models applied to cybersecurity (for a review, see Veksler et al., 2018; Veksler et al., 2020). Veksler et al. (2020) argue that such cognitive models can used to predict the behaviour of attackers or computer system users. For example, Sandouka et al. (2009) used neural network models to detect social engineering attacks. The model was applied to phone conversation data, which include logs of phone calls. Each log includes date, time, where the call originated and terminated, and details of the conversation (Hoeschele, 2006). The model was used to analyse the text and detect any intrusions or social engineering attempts. Furthermore, Maqbool et al. (2020) used cognitive modeling and found that an excessive reliance on recency and frequency are related to cyber-attacks. However, future work should use computational models to better understand the relationship between cognitive processes and cybersecurity behaviours.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

FUNDING

AM acknowledges funding from Socially Engineered Payment Diversion Fraud granted by the NSW Cybersecurity Innovation Node.

REFERENCES

Aggarwal, P., Frédéric, M., Gonzalez, M. C., and Dutt, V. (2018). Understanding cyber situational awareness in a cyber security game involving recommendation. *Int. J. Cyber Situat. Aware.* 3, 11–38. doi: 10.22619/ijcsa.2018.100118

Akhawe, D., and Felt, A. P. (2013). "Alice in warningland: a large-scale field study of browser security warning effectiveness," in *Proceedings of the 22nd USENIX Security Symposium*, Washington, DC.

Alonso, C., and Romero, E. (2017). Aggressors and victims in bullying and cyberbullying: a study of personality profiles using the five-factor model. *Span. J. Psychol.* 20:e76.

Altintas, E., Karaca, Y., Moustafa, A. A., and El Haj, M. (2020). Effect of best possible self intervention on situational motivation and commitment in academic context. *Learn. Motiv.* 69:101599. doi: 10.1016/j.lmot.2019.101599

Anderson, B. B., Kirwan, C. B., Jenkins, J. L., and Eargle, D. (2015). "How polymorphic warnings reduce habituation in the brain—insights from an fmri study," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems CHI, Crossings*, Seoul.

Anderson, C. L., and Agarwal, R. (2010). Practicing safe computing: a multimethod empirical examination of home computer user security behavioral intentions. *MIS Q.* 34, 613–643. doi: 10.2307/25750694

Andrade, R. O., and Yoo, S. G. (2019). Cognitive security: a comprehensive study of cognitive science in cybersecurity. *J. Inf. Secur. Appl.* 48:102352. doi: 10.1016/j.jisa.2019.06.008

Bailey, P. E., Slessor, G., Rieger, M., Rendell, P. G., Moustafa, A. A., and Ruffman, T. (2015). Trust and trustworthiness in young and older adults. *Psychol. Aging* 30, 977–986. doi: 10.1037/a0039736

Baillon, A., de Bruin, J., Emirmahmutoglu, A., van de Veer, E., and van Dijk, B. (2019). Informing, simulating experience, or both: a field experiment on phishing risks. *PLoS One* 14:e0224216. doi: 10.1371/journal.pone.0224216

Barford, P. (2010). "Cyber SA: situational awareness for cyber defense," in *Cyber Situational Awareness.* Advances in Information Security, Vol. 46, eds P. Liu, S. Jajodia, V. Swarup, and C. Wang (Boston, MA: Springer).

Benson, V., and Mcalaney, J. (2020). *Cyber Influence and Cognitive Threats*. Cambridge, MA: Academic Press.

Blais, A. R., and Weber, E. U. (2006). A domain-specific risk-taking (dospert) scale for adult populations. *Judgm. Decis. Mak.* 1, 33–47.

Bodi, N., Keri, S., Nagy, H., Moustafa, A., Myers, C. E., Daw, N., et al. (2009). Reward-learning and the novelty-seeking personality: a between- and within-subjects study of the effects of dopamine agonists on young Parkinson's patients. *Brain* 132(Pt 9), 2385–2395. doi: 10.1093/brain/awp094

Bowen, B. M., Devarajan, R., and Stolfo, S. (2014). "Measuring the human factor of cyber security," in *Proceedings of the 2011 IEEE International Conference on Technologies for Homeland Security (HST)*, Waltham, MA.

Boyce, M. W., Duma, K. M., Hettinger, L. J., Malone, T. B., Wilson, D. P., and Lockett-Reynolds, J. (2011). "Human performance in cybersecurity: a research agenda ," in *Proceedings of the Human Factors and Ergonomics Society 55th Annual Meeting*.

Brase, G. L., Vasserman, E. Y., and Hsu, W. (2017). Do different mental models influence cybersecurity behavior? Evaluations via statistical reasoning performance. *Front. Psychol.* 8:1929. doi: 10.3389/fpsyg.2017.01929

Bravo-Lillo, C., Komanduri, S., Cranor, L., Reeder, R., Sleeper, M., Downs, J., et al. (2013). "Your attention please: designing security-decision UIs to make genuine risks harder to ignore," in *Proceedings of the Symposium on Usable Privacy and Security (SOUPS)*, Newcastle.

Bromberg, U., Lobatcheva, M., and Peters, J. (2017). Episodic future thinking reduces temporal discounting in healthy adolescents. *PLoS One* 12:e0188079. doi: 10.1371/journal.pone.0188079

Bromberg, U., Wiehler, A., and Peters, J. (2015). Episodic future thinking is related to impulsive decision making in healthy adolescents. *Child. Dev.* 86, 1458–1468. doi: 10.1111/cdev.12390

Brustoloni, J. C., and Villamarín-Salomón, R. (2007). "Improving security decisions with polymorphic and audited dialogs," in *Proceedings of the SOUPS '07: 3rd Symposium on Usable Privacy and Security*, New York, NY, 76–85. doi: 10.1145/1280680.1280691

Cacioppo, J. T., Petty, R. E., and Feng Kao, C. (1984). The efficient assessment of need for cognition. *J. Pers. Assess.* 48, 306–307. doi: 10.1207/s15327752jpa4803_13

Calic, D., Pattinson, M., and Parsons, K. (2016). "Naive and accidental behaviours that compromise information security: what the experts think," in *Proceedings of the 10th International Symposium of Human Aspects of Information Security and Assurance*, eds N. L. Clarke and S. M. Furnell (Frankfurt: HAISA).

Chan, M., Woon, I. M. Y., and Kankanhalli, A. (2005). Perceptions of information security at the workplace: linking information security climate to compliant behavior. *J. Inf. Privacy Secur.* 1, 18–41. doi: 10.1080/15536548.2005.10855772

Curtis, S., Basak, A., Carre, J., Bošanskı, B., Èernı, J., Ben-Asher, N., et al. (2021). The Dark Triad and strategic resource control in a competitive computer game. *Pers. Individ. Diff.* 168:110343. doi: 10.1016/j.paid.2020.110343

Curtis, S. R., Rajivan, P., Jones, D. N., and Gonzalez, C. (2018). Phishing attempts among the dark triad: patterns of attack and vulnerability. *Comput. Hum. Behav.* 87, 174–182. doi: 10.1016/j.chb.2018.05.037

Cybenko, G., Giani, A., and Thompson, P. (2002). Cognitive hacking: a battle for the mind. *Computer* 35, 50–56. doi: 10.1109/mc.2002.1023788

D'Amico, A., Whitley, K., and Tesone, D. (2005). "Achieving cyber defense situational awareness: a cognitive task analysis of information assurance analysts," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Los Angeles, CA.

Dawson, J., and Thomson, R. (2018). The future cybersecurity workforce: going beyond technical skills for successful cyber performance. *Front. Psychol.* 9:744. doi: 10.3389/fpsyg.2018.00744

Diaz, A., Sherman, A. T., and Joshi, A. (2018). Phishing in an academic community: a study of user susceptibility and behavior. *arXiv* [Preprint] arXiv:1811.06078,

Dutt, V., Ahn, Y., and Gonzalez, C. (2013). Cyber situation awareness: modeling detection of cyber attacks with instance-based learning theory. *Hum. Factors* 55, 605–618. doi: 10.1177/0018720812464045

Egelman, S., and Peer, E. (2015). Scaling the security wall developing a security behavior intentions scale (SEBIS). *Paper Presented at the Security Feedback & Warnings CHI*, Seoul.

Eskritt, M., Doucette, J., and Robitaille, L. (2014). Does future-oriented thinking predict adolescent decision making? *J. Genet. Psychol.* 175, 163–179. doi: 10.1080/00221325.2013.875886

Frank, M. J., Moustafa, A. A., Haughey, H. M., Curran, T., and Hutchison, K. E. (2007). Genetic triple dissociation reveals multiple roles for dopamine in reinforcement learning. *Proc. Natl. Acad. Sci. U.S.A.* 104, 16311–16316. doi: 10.1073/pnas.0706111104

Fraunholz, D., Anton, S. D., Lipps, C., Reti, D., Krohmer, D., Pohl, F., et al. (2019). Demystifying deception technology: a survey. *arXiv* [Preprint] arXiv:1804.06196,

Furnell, S., and Clarke, C. (2012). Power to the people? The evolving recognition of human aspects of security. *Comput. Secur.* 31, 983–988. doi: 10.1016/j.cose.2012.08.004

Garrett, N., and Sharot, T. (2017). Optimistic update bias holds firm: three tests of robustness following Shah et al. *Conscious Cogn.* 50, 12–22. doi: 10.1016/j.concog.2016.10.013

Goodboy, A., and Martin, M. (2015). The personality profile of a cyberbully: examining the dark triad. *Comput. Hum. Behav.* 49, 1–4. doi: 10.1016/j.chb.2015.02.052

Greenwald, S. J., Olthoff, K. G., Raskin, V., and Ruch, W. (2004). The user non-acceptance paradigm: INFOSEC's dirty little secret. *Paper Presented at the New Security Paradigms Workshop*, New York, NY.

Guo, K. H., Yuan, Y., Archer, N. P., and Connelly, C. E. (2011). Understanding nonmalicious security violations in the workplace: a composite behavior model. *J. Manag. Inf. Syst.* 28, 203–236. doi: 10.2753/mis0742-1222280208

Gutzwiller, R. S., Fugate, S., Sawyer, B. D., and Hancock, P. A. (2015). The human factors of cyber network defense. *Paper presented at the In Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Los Angeles, CA.

Hadlington, L. (2017). Human factors in cybersecurity; examining the link between Internet addiction, impulsivity, attitudes towards cybersecurity, and risky cybersecurity behaviours. *Heliyon* 3:e00346. doi: 10.1016/j.heliyon.2017.e00346

Hadlington, L., and Murphy, K. (2018). Is media multitasking good for cybersecurity? exploring the relationship between media multitasking and everyday cognitive failures on self-reported risky cybersecurity behaviors. *Cyberpsychol. Behav. Soc. Netw.* 21, 168–172. doi: 10.1089/cyber.2017.0524

Hakim, Z. M., Ebner, N. C., Oliveira, D. S., Getz, S. J., Levin, B. E., Lin, T., et al. (2020).). The phishing email suspicion test (PEST) a lab-based task for evaluating the cognitive mechanisms of phishing detection. *Behav. Res. Methods* doi: 10.3758/s13428-020-01495-0 [Epub ahead of print].

Halevi, T., Lewis, J., and Memon, N. (2013). "A pilot study of cyber security and privacy related behaviour and personality traits," in *Proceedings of the WWW '13 Companion: 22nd International Conference on World Wide Web*, Rio de Janeiro.

Hamill, R. F., and Deckro, J. M. K. (2005). Evaluating information assurance strategies. *Decis. Support Syst.* 39, 463–484. doi: 10.1016/j.dss.2003.11.004

Harrison, A., Summers, J., and Mennecke, B. (2018). The effects of the dark triad on unethical behavior. *J. Bus. Ethics* 153, 53–77. doi: 10.1007/s10551-016-3368-3

Hazari, S., Hargrave, W., and Clenney, B. (2009). An empirical investigation of factors influencing information security behavior. *J. Inf. Privacy Secur.* 4, 3–20. doi: 10.1080/2333696x.2008.10855849

Henshel, D., Cains, M. G., Hoffman, B., and Kelley, T. (2015). Trust as a human factor in holistic cyber security risk assessment. *Proc. Manuf.* 3, 1117–1124. doi: 10.1016/j.promfg.2015.07.186

Herath, T., and Rao, H. R. (2009). Protection motivation and deterrence: a framework for security policy compliance in organisations. *Eur. J. Inf. Syst.* 18, 106–125. doi: 10.1057/ejis.2009.6

Herzallah, M. M., Moustafa, A. A., Natsheh, J. Y., Abdellatif, S. M., Taha, M. B., Tayem, Y. I., et al. (2013). Learning from negative feedback in patients with major depressive disorder is attenuated by SSRI antidepressants. *Front. Integr. Neurosci.* 7:67. doi: 10.3389/fnint.2013.00067

Hoeschele, M. (2006). *Detecting Social Engineering*. CERIAS Tech Report 2006-15. Ph.D. Thesis. West Lafayette, IN: Purdue University.

Hu, Q., West, R., and Smarandescu, L. (2015). The role of self-control in information security violations: insights from a cognitive neuroscience perspective. *J. Manag. Inf. Syst.* 31, 6–48. doi: 10.1080/07421222.2014.1001255

Ifinedo, P. (2014). Information systems security policy compliance: an empirical study of the effects of socialisation, influence, and cognition. *Inf. Manag.* 51, 69–79. doi: 10.1016/j.im.2013.10.001

Jacobs, N., Goossens, L., Dehue, F., Völlink, T., and Lechner, L. (2015). Dutch cyberbullying victims' experiences, perceptions, attitudes and motivations related to (coping with) cyberbullying: focus group interviews. *Societies* 5, 43–64. doi: 10.3390/soc5010043

Jagatic, T. N., Johnson, N. A., Jakobsson, M., and Menczer, F. (2007). Social phishing. *Commun. ACM* 50, 94–100.

Jakobsson, M., and Ratkiewicz, J. (2006). "Designing ethical phishing experiments: a study of (ROT13) rOnl query features," in *Proceedings of the 15th International Conference on World Wide Web Feature*, Scotland

Joireman, J., Shaffer, M. J., Balliet, D., and Strathman, A. (2012). Promotion orientation explains why future-oriented people exercise and eat healthy evidence from the two-factor consideration of future consequences-14 scale. *Pers. Soc. Psychol. Bull.* 38, 1272–1287. doi: 10.1177/0146167212449362

Jones, A., and Colwill, C. (2008). "Dealing with the Malicious Insider," in *Proceedings of the 6th Australian Information Security Management Conference*, (Perth,WA: Edith Cowan University).

Kar, K., Moustafa, A. A., Myers, C. E., and Gluck, M. A. (2010). "Using an animal learning model of the hippocampus to simulate human fMRI data," in *Proceedings of the 2010 IEEE 36th Annual Northeast Bioengineering Conference (NEBEC)*, New York, NY.

Keller, U., Strobel, A., Wollschläger, R., Greiff, S., Martin, R., Vainikainen, M., et al. (2019). A need for cognition scale for children and adolescents: structural analysis and measurement invariance. *Eur. J. Psychol. Assess.* 35, 137–149. doi: 10.1027/1015-5759/a000370

King, Z. M., Henshel, D. S., Flora, L., Cains, M. G., Hoffman, B., and Sample, C. (2018). Characterizing and measuring maliciousness for cybersecurity risk assessment. *Front. Psychol.* 9:39. doi: 10.3389/fpsyg.2018.00039

Knott, B. A., Mancuso, V. F., Bennett, K., Finomore, V., McNeese, M., and McKneely, J. A. (2013). "Human factors in cyber warfare: alternative perspectives," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Los Angeles, CA.

Landwehr, C. (1981). Formal models of computer security. *Comput. Surv.* 13, 247–278. doi: 10.1145/356850.356852

Lee, Y., and Kozar, K. A. (2005). Investigating factors affecting the adoption of anti-spyware systems. *Commun. ACM* 48, 72–77. doi: 10.1145/1076211.1076243

Lin, Y., Durbin, J. M., and Rancer, A. S. (2016). Math anxiety, need for cognition, and learning strategies in quantitative communication research methods courses. *Commun. Q.* 64, 390–409. doi: 10.1080/01463373.2015.1103294

Maasberg, M., Van Slyke, C., Ellis, S., and Beebe, N. (2020). The dark triad and insider threats in cyber security. *Commun. ACM* 63, 64–80. doi: 10.1145/3408864

Mancuso, M., Christensen, J. C., Cowley, J., and Finomore, V. (2014). "Human factors in cyber warfare II: emerging perspectives," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, (Thousand Oaks, CA: SAGE Publications), 58.

Maqbool, Z., Aggarwal, P., Pammi, V. S. C., and Dutt, V. (2020). Cyber security: effects of penalizing defenders in cyber-security games via experimentation and computational modeling. *Front. Psychol.* 11:11. doi: 10.3389/fpsyg.2020.00011

Maurushat, A. (2010). Hackers, Fraudsters and Botnets: Tackling the Problem of Cyber Crime – The Report of the Inquiry into Cyber Crime Invited Submission to the House of Representatives Standing Committee on Communications, Parliament of Australia. Available online at: http://aph.gov.au/house/committee/coms/cybercrime/report/full_report.pdf

May, A. C., Aupperle, R. L., and Stewart, J. L. (2020). Dark times: the role of negative reinforcement in methamphetamine addiction. *Front. Psychiatry* 11:114. doi: 10.3389/fpsyt.2020.00114

McAlaney, J., Taylor, J., and Faily, S. (2015). The social psychology of cybersecurity. *Paper presented at the 1st International Conference on Cyber Security for Sustainable Society. Coventry*.

McBride, M., Carter, L., and Warkentin, M. (2012). Exploring the role of individual employee characteristics and personality on employee compliance with cybersecurity policies. *RTI Int. Inst. Homel. Secur. Solut.* 5:1. doi: 10.1016/j.paid.2019.05.040

Mishra, S., and Dhillon, G. (2006). "Information systems security governance research: a behavioral perspective," in *Proceedings of the 1st Annual Symposium on Information Assurance, academic track of the 9th Annual 2006 NYS Cyber Security Conference*, New York, NY.

Mohebzada, J., El Zarka, A., BHojani, A. H., and Darwish, A. (2012). "Phishing in a university community: Two large scale phishing experiments," in *Proceedings of the Innovations in Information Technology (IIT), International Conference*, (Piscataway, NJ: IEEE), 249–254.

Moustafa, A. A., Cohen, M. X., Sherman, S. J., and Frank, M. J. (2008). A role for dopamine in temporal decision making and reward maximization in parkinsonism. *J. Neurosci.* 28, 12294–12304. doi: 10.1523/jneurosci.3116-08.2008

Moustafa, A. A., Keri, S., Herzallah, M. M., Myers, C. E., and Gluck, M. A. (2010). A neural model of hippocampal-striatal interactions in associative learning and transfer generalization in various neurological and psychiatric patients. *Brain Cogn.* 74, 132–144. doi: 10.1016/j.bandc.2010.07.013

Moustafa, A. A., Keri, S., Polner, B., and White, C. (2017). Drift diffusion model of reward and punishment learning in rare alpha-synuclein gene carriers. *J. Neurogenet.* 31, 17–22. doi: 10.1080/01677063.2017.1301939

Moustafa, A. A., Keri, S., Somlai, Z., Balsdon, T., Frydecka, D., Misiak, B., et al. (2015). Drift diffusion model of reward and punishment learning in schizophrenia: modeling and experimental data. *Behav. Brain Res.* 291, 147–154. doi: 10.1016/j.bbr.2015.05.024

Moustafa, A. A., Krishna, R., Eissa, A. M., and Hewedi, D. H. (2013). Factors underlying probabilistic and deterministic stimulus-response learning performance in medicated and unmedicated patients with Parkinson's disease. *Neuropsychology* 27, 498–510. doi: 10.1037/a0032757

Moustafa, A. A., Morris, A. N., and ElHaj, M. (2018a). A review on future episodic thinking in mood and anxiety disorders. *Rev. Neurosci.* 30, 85–94. doi: 10.1515/revneuro-2017-0055

Moustafa, A. A., Morris, A. N., Nandrino, J. L., Misiak, B., Szewczuk-Boguslawska, M., Frydecka, D., et al. (2018b). Not all drugs are created equal: impaired future thinking in opiate, but not alcohol, users. *Exp. Brain. Res.* 236, 2971–2981. doi: 10.1007/s00221-018-5355-7

Moustafa, A. A., Myers, C. E., and Gluck, M. A. (2009). A neurocomputational model of classical conditioning phenomena: a putative role for the hippocampal region in associative learning. *Brain Res.* 1276, 180–195. doi: 10.1016/j.brainres.2009.04.020

Moutsiana, C., Garrett, N., Clarke, R. C., Lotto, R. B., Blakemore, S. J., and Sharot, T. (2013). Human development of the ability to learn from bad news. *Proc. Natl. Acad. Sci. U.S.A.* 110, 16396–16401. doi: 10.1073/pnas.1305631110

Myers, C. E., Sheynin, J., Balsdon, T., Luzardo, A., Beck, K. D., Hogarth, L., et al. (2016). Probabilistic reward- and punishment-based learning in opioid addiction: experimental and computational data. *Behav. Brain Res.* 296, 240–248. doi: 10.1016/j.bbr.2015.09.018

Nobles, C. (2018). Botching human factors in cybersecurity in business organizations. *Holistica* 9, 71–88. doi: 10.2478/hjbpa-2018-0024

Parsons, K., Calic, D., Pattinson, M., Butavicius, M., McCormaca, A., and Zwaans, T. (2017). The human aspects of information security questionnaire (HAIS-Q): two further validation studies. *Comput. Secur.* 55, 40–51. doi: 10.1016/j.cose.2017.01.004

Patton, J. H., Stanford, M. S., and Barratt, E. S. (1995). Factor structure of the barratt impulsiveness scale. *J. Clin. Psychol.* 51, 768–774. doi: 10.1002/1097-4679(199511)51:6<768::aid-jclp2270510607>3.0.co;2-1

Paulhus, D., and Williams, K. (2002). The dark triad of personality: narcissism, machiavellianism, and psychopathy. *J. Res. Pers.* 36, 556–563. doi: 10.1016/s0092-6566(02)00505-6

Peckham, A. D., and Johnson, S. L. (2018). Cognitive control training for emotion-related impulsivity. *Behav. Res. Ther.* 105, 17–26. doi: 10.1016/j.brat.2018.03.009

Piray, P., Zeighami, Y., Bahrami, F., Eissa, A. M., Hewedi, D. H., and Moustafa, A. A. (2014). Impulse control disorders in Parkinson's disease are associated with dysfunction in stimulus valuation but not action valuation. *J. Neurosci.* 34, 7814–7824. doi: 10.1523/jneurosci.4063-13.2014

Rajivan, P., Aharonov-Majar, E., and Gonzalez, C. (2020). Update now or later? Effects of experience, cost, and risk preference on update decisions. *J. Cyber Secur.* 6:tyaa002.

Rajivan, P., and Gonzalez, C. (2018). Creative persuasion: a study on adversarial behaviors and strategies in phishing attacks. *Front. Psychol.* 9:135. doi: 10.3389/fpsyg.2018.00135

Rankin, C. H., Abrams, T., Barry, R. J., Bhatnagar, S., Clayton, D. F., Colombo, J., et al. (2009). Habituation revisited: an updated and revised description of the behavioral characteristics of habituation. *Neurobiol. Learn. Mem.* 92, 135–138. doi: 10.1016/j.nlm.2008.09.012

Regier, P. S., and Redish, A. D. (2015). Contingency management and deliberative decision-making processes. *Front. Psychiatry* 6:76. doi: 10.3389/fpsyt.2015.00076

Rodriguez-Enriquez, M., Bennasar-Veny, M., Leiva, A., Garaigordobil, M., and Yanez, A. M. (2019). Cybervictimization among secondary students: social networking time, personality traits and parental education. *BMC Public Health* 19:1499. doi: 10.1186/s12889-019-7876-9

Rosenbaum, G. M., Botdorf, M. A., Patrianakos, J. L., Steinberg, L., and Chein, J. M. (2017). Working memory training in adolescents decreases laboratory risk taking in the presence of peers. *J. Cogn. Enhanc.* 1, 513–525. doi: 10.1007/s41465-017-0045-0

Sadkhan, S. B. (2019). Cognition and the future of information security. *Paper presented at the 2019 International Conference on Advanced Science and Engineering (ICOASE).*

Saleme, D., and Moustafa, A. A. (2020). "The multifaceted nature of risk-taking in drug addiction," in *Cognitive, Clinical, and Neural Aspects of Drug Addiction*, ed. A. A. Moustafa (Amsterdam: Elsevier).

Saleme, D. M., Kluwe-Schiavon, B., Soliman, A., Misiak, B., Frydecka, D., and Moustafa, A. A. (2018). Factors underlying risk taking in heroin-dependent individuals: Feedback processing and environmental contingencies. *Behav. Brain Res.* 350, 23–30. doi: 10.1016/j.bbr.2018.04.052

Sandouka, H., Cullen, A., and Mann, I. (2009). Social engineering detection using neural networks. *Paper Presented at the International Conference on CyberWorlds.*

Sarmany-Schuller, I. (1999). Procrastination, need for cognition and sensation seeking. *Stud. Psychol.* 41, 73–85.

Sasse, M. A., Brostoff, S., and Weirich, D. (2004). Transforming the weakest link – a human/ computer interaction approach to usable and effective security. *BT Technol. J.* 19, 122–131.

Schechter, S., Dhamija, R., Ozment, A., and Fischer, I. (2007). "The emperor's new security indicators," in *Proceedings of the IEEE Symposium on Security and Privacy*, Berkeley, CA.

Schneier, B. (2004). *Secrets and Lies: Digital Security in a Networked World.* Hoboken, NJ: Wiley.

Scott, S. G., and Bruce, R. A. (1995). Decision-making style: the development and assessment of a new measure. *Educ. Psychol. meas.* 55, 818–831. doi: 10.1177/0013164495055005017

Sharot, T. (2011). The optimism bias. *Curr. Biol.* 21, R941–R945. doi: 10.1016/j.cub.2011.10.030

Shropshire, J., Warkentin, M., Johnston, A. C., and Schmidt, M. B. (2006). Personality and IT security: an application of the five-factor model. *Paper presented at the Connecting the Americas, 12th Americas Conference on Information Systems*, (Acapulco: AMCIS).

Sidman, M. (2006). The distinction between positive and negative reinforcement: some additional considerations. *Behav. Anal.* 29, 135–139. doi: 10.1007/bf03392126

Smith, S. W. (2003). Humans in the loop human–computer interaction and security. *IEEE Comput. Soc.* 1, 75–79. doi: 10.1109/msecp.2003.1203228

Stanton, J. M., Stam, J. R., Mastrangelo, P. M., and Jolton, J. A. (2005). Analysis of end user security behaviors. *Comput. Secur.* 24, 124–133. doi: 10.1016/j.cose.2004.07.001

Thompson, P. (2004). Cognitive hacking and intelligence and security informatics. *Proc. SPIE* 5423, 142–151.

Tversky, A., and Kahneman, D. (1986). Rational choice and the framing of decisions. *J. Bus.* 59, 251–278.

Tyworth, M., Giacobe, N. A., Mancuso, V. F., McNeese, M. D., and Hall, D. L. (2013). A human-in-the-loop approach to understanding situation awareness in cyber defence analysis. *EAI Endorsed Trans. Secur. Safe.* 13:e6. doi: 10.4108/trans.sesa.01-06.2013.e6

Veksler, V. D., Buchler, N., Hoffman, B. E., Cassenti, D. N., Sample, C., and Sugrim, S. (2018). Simulations in cyber-security: a review of cognitive modeling of network attackers, defenders, and users. *Front. Psychol.* 9:691. doi: 10.3389/fpsyg.2018.00691

Veksler, V. D., Buchler, N., LaFleur, C. G., Yu, M. S., Lebiere, C., and Gonzalez, C. (2020). Cognitive models in cybersecurity: learning from expert analysts and predicting attacker behavior. *Front. Psychol.* 11:1049. doi: 10.3389/fpsyg.2020.01049

Vroom, C., and von Solms, R. (2004). Towards information security behavioural compliance. *Comput. Secur.* 23, 191–198. doi: 10.1016/j.cose.2004.01.012

West, R. (2008). The psychology of security: why do good users make bad decisions. *Commun. ACM* 51, 34–40.

Whitty, M., Doodson, J., Creese, S., and Hodges, D. (2015). Individual differences in cyber security behaviors: an examination of who is sharing passwords. *Cyberpsychol. Behav. Soc. Netw.* 18, 3–7. doi: 10.1089/cyber.2014.0179

Wiederhold, B. K. (2014). The role of psychology in enhancing cybersecurity. *Cyberpsychol. Behav. Soc. Netw.* 17, 131–132. doi: 10.1089/cyber.2014.1502

Wogalter, M. S. (2006). "Communication-human information processing (C-HIP) model," in *Handbook of Warnings*, ed. M. S. Wogalter (Mahwah, N.J: Erlbaum), 51–61.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read
for greatest visibility
and readership

FAST PUBLICATION
Around 90 days
from submission
to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative,
and constructive
peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers
acknowledged by name
on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data
and methods to enhance
research reproducibility

DIGITAL PUBLISHING
Articles designed
for optimal readership
across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics
track visibility across
digital media

EXTENSIVE PROMOTION
Marketing
and promotion
of impactful research

LOOP RESEARCH NETWORK
Our network
increases your
article's readership