


Persuasion Meets AI: Ethical Considerations for the Design of Social Engineering Countermeasures

Nicolás E. Díaz Ferreyra¹ ^a, Esma Aïmeur², Hicham Hage³, Maritta Heisel¹
and Catherine García van Hoogstraten⁴

¹Department of Computer Science and Applied Cognitive Science, University of Duisburg-Essen, Germany

²Department of Computer Science and Operations Research (DIRO), University of Montréal, Canada

³Computer Science Department, Notre Dame University - Louaize, Lebanon

⁴Booking.com, The Netherlands

Keywords: Social Engineering, Digital Nudging, Privacy Awareness, Risk Management, Self-disclosure, AI, Ethics.

Abstract: Privacy in Social Network Sites (SNSs) like Facebook or Instagram is closely related to people's self-disclosure decisions and their ability to foresee the consequences of sharing personal information with large and diverse audiences. Nonetheless, online privacy decisions are often based on spurious risk judgements that make people liable to reveal sensitive data to untrusted recipients and become victims of social engineering attacks. Artificial Intelligence (AI) in combination with persuasive mechanisms like nudging is a promising approach for promoting preventative privacy behaviour among the users of SNSs. Nevertheless, combining behavioural interventions with high levels of personalization can be a potential threat to people's agency and autonomy even when applied to the design of social engineering countermeasures. This paper elaborates on the ethical challenges that nudging mechanisms can introduce to the development of AI-based countermeasures, particularly to those addressing unsafe self-disclosure practices in SNSs. Overall, it endorses the elaboration of personalized risk awareness solutions as i) an ethical approach to counteract social engineering, and ii) as an effective means for promoting reflective privacy decisions.

1 INTRODUCTION

Social Network Sites (SNSs) like Instagram and Twitter have changed radically the way people create and maintain interpersonal relationships (Penni, 2017). One of the major attractiveness of such platforms is their broadcasting affordances which allow users to connect seamlessly with large and diverse audiences within a few seconds. However, while these platforms effectively contribute to maximizing people's social capital, they also introduce major challenges to their privacy. Particularly, users of SNSs are highly exposed to social engineering attacks since they often share personal information with people regardless of their doubtful trustworthiness (Wang et al., 2011; boyd, 2010).

Online deception is an attack vector that is frequently used by social engineers to approach and manipulate users of SNSs (Tsikerdekis and Zeadally,

2014). For instance, deceivers often impersonate trustworthy entities using fake profiles to gain their victims' trust, and persuade them to reveal sensitive information (e.g. their log-in credentials) or perform hazardous actions that would compromise their security (e.g. installing Malware) (Hage et al., 2020). Hence, counteracting social engineering attacks relies (to a large extent) on the users' capacity of foreseeing the potential negative consequences of their actions and modify their behaviour, accordingly (Aïmeur et al., 2019). However, this is difficult for average users who lack the knowledge and skills necessary to ensure the protection of their privacy (Masur, 2019). Moreover, people -in general- regret having shared their personal information only after being victims of a social engineering attack (Wang et al., 2011).

Privacy scholars have proposed a wide range of Artificial-Intelligence-Based (AI-based) approaches that aim at generating awareness among people (De

^a  <https://orcid.org/0000-0001-6304-771X>

and Le Métayer, 2018; Petkos et al., 2015; Díaz Ferreyra et al., 2020; Briscoe et al., 2014). Particularly, AI in combination with persuasive mechanisms has gained popularity due to their capacity for nudging users' behaviour towards safer privacy practices (Acquisti et al., 2017). However, these technologies are often looked askance since there is a fine line between persuasion, coercion, and manipulation (Renaud and Zimmermann, 2018). For instance, users might be encouraged to enable their phone's location services with the excuse of increasing their safety when, in fact, the main objective is monitoring their movements. Hence, ethical principles must be well defined and followed for safeguarding people's agency, autonomy, and welfare.

This work elaborates on the ethical challenges associated with the use of persuasion in the design of AI-based countermeasures (i.e. technical solutions for preventing unsafe self-disclosure practices). Particularly, it analyses the different factors that influence online privacy decisions and the importance of behavioural interventions for promoting preventative privacy practices in SNSs. Furthermore, it elaborates on the ethical issues that the use of persuasive means may introduce when used in combination with AI technologies. Based on our findings, we endorse the elaboration of personalized risk awareness mechanisms as an ethical approach to social engineering countermeasures. In line with this, challenges for regulating the development and impact of such countermeasures are evaluated and summarized.

The rest of this paper is organized as follows. Section 2 discusses the particular challenges that SNSs introduce in terms of privacy decision-making. Section 3 elaborates on the use of privacy nudges in combination with AI for designing effective social engineering countermeasures. Moreover, ethical challenges related to the use of persuasion in cybersecurity are presented and illustrated in this section. Next, Section 4 analyses the role of risk cues in users' privacy decisions and their importance for the design of preventative technologies. Finally, the conclusions of this work are presented in Section 5.

2 ONLINE DECEPTION IN SOCIAL MEDIA

Nowadays, SNSs offer a wide range of affordances (e.g. instant messaging, posts, or stories) which allow people to create and exchange media content with large and diverse audiences. In such a context, privacy as a human practice (i.e. as a decision-making process) acquires high importance since individuals

are prone to disclose large amounts of private information inside these platforms (boyd, 2010). Consequently, preserving users' contextual integrity depends to a wide extent on their individual behaviour, and not so much on the security mechanisms of the platform (e.g. firewalls or communication protocols) (Albladi and Weir, 2016). In general, disclosing personal information to others is key for the development and strengthening of social relationships, as it directly contributes to building trust and credibility among individuals. However, unlike in the real world, people in SNSs tend to reveal their personal data prematurely without reflecting much on the potential negative effects (Aïmeur et al., 2018). On one hand, such spurious behaviour can be grounded on users' ignorance and overconfidence (Howah and Chugh, 2019). Nevertheless, people often rely on lax privacy settings and assume their online peers as trusted, which increases significantly the chances of being victims of a malicious user. Therefore, individuals are prone to experience unwanted incidents like cyber-bullying, reputation damage, or identity theft after sharing their personal information in online platforms (Wang et al., 2013).

Overall, SNSs have become a gateway for accessing large amounts of personal information and, consequently, a target for social engineering attacks. On one hand, this is because people are more liable to reveal personal information online than in a traditional offline context. However, there is also a growing trend in cyber-attacks to focus more on human vulnerabilities instead on flaws in software or hardware (Krombholz et al., 2015). Moreover, it is estimated that around 3% of Malware attacks exploit technical lapses while the remaining 97% target the users using social engineering¹. Basically, social engineers employ online deception as a strategy to gain trust and manipulate their victims. Particularly, "deceivers hide their harmful intentions and mislead other users to reveal their credentials (i.e. accounts and passwords) or perform hazardous actions (e.g. install Malware)" (Aïmeur et al., 2019). For instance, they often approach users through fake SNSs accounts and instigate them to install malicious software on their computers. For this, deceivers exploit users' motivations and cognitive biases such as altruism or moral gain in combination with incentive strategies to mislead them, accordingly (Bullée et al., 2018). Particularly, the use of fake links to cash prizes or fake surveys on behalf of trustworthy entities can serve as incentives

¹"Estimates of the number of social engineering based cyber-attacks into private or government organization," DOGANA H2020 Project. Accessed July 24, 2020. <https://bit.ly/2k5VKmP>

and, thereby, as deceptive means.

3 AI-BASED COUNTERMEASURES

In general, people struggle to regulate the amount of information they share as they seek the right balance between self-disclosure gratifications and privacy risks. Moreover, an objective evaluation of such risks demands a high cognitive effort which is often affected by personal characteristics, emotions, or missing knowledge (Krämer and Schäwel, 2020). Hence, there is a call for technological countermeasures that support users in conducting a more accurate privacy calculus and incentivize the adoption of preventative behaviour. In this section, we discuss the role of AI in the design of such countermeasures especially in combination with persuasive technologies like digital nudges. Furthermore, ethical guidelines for the application of these technologies are presented and analysed.

3.1 Privacy Nudges

The use of persuasion in social computing applications like blogs, wiki, and recently SNSs has caught the interest of researchers across a wide range of disciplines including computer science and cognitive psychology (Vassileva, 2012). Additionally, the field of behavioural economics has contributed largely to this topic and nourished several principles of user engagement such as gamification or incentive mechanisms for promoting behavioural change (Hamari and Koivisto, 2013). Most recently, the nudge theory and its application for privacy and security purposes have been closely explored and documented within the literature (Acquisti et al., 2017). Originally, the term nudge was coined by the Nobel prize winners Richard Thaler and Cass Sunstein and refers to the introduction of small changes in a choice architecture (i.e. the context within which decisions are made) with the purpose of encouraging a certain user behaviour (Weinmann et al., 2016). Among its many applications, the nudge concept has been applied in the design of preventative technologies with the aim of guiding users towards safer privacy decisions. For example, (Wang et al., 2013) designed three nudges for Facebook users consisting of (i) introducing a 30 seconds delay before a message is posted, (ii) displaying visual cues related to the post’s audience, and (iii) showing information about the sentiment of the post. These nudges come into play when users are about

to post a message on Facebook allowing them to reconsider their disclosures and reflect on the potential privacy consequences. Moreover, nudges have also been designed, developed, and applied for security purposes. This is the case of password meters used to promote stronger passwords (Egelman et al., 2013) or the incorporation of visual cues inside Wi-Fi scanners to encourage the use of secure networks (Turland et al., 2015).

3.2 The Role of AI

In general, the instances of privacy nudges described in the current literature rely on a “one-size-fits-all” persuasive design. That is, the same behavioural intervention is applied to diverse individuals without acknowledging the personal characteristics or differences among them (Warberg et al., 2019). However, there is an increasing demand for personalized nudges that address nuances in users’ privacy goals and regulate their interventions, accordingly (Peer et al., 2020; Barev and Janson, 2019).

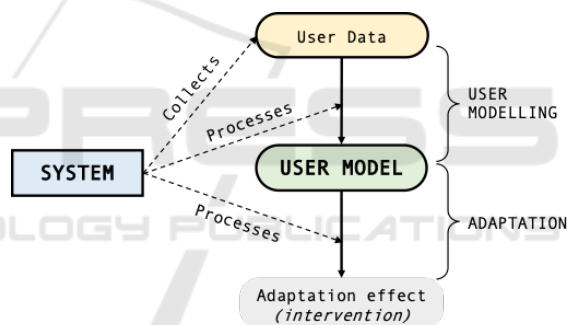


Figure 1: The user modelling-adaptation loop (De Bra, 2017).

In essence, the idea of personalized nudges inherently encloses the application of AI techniques and methods for understanding and anticipating the privacy needs of each particular user. For this, it is necessary to define what is commonly known as the “user model” of the system. That is, a set of adaptation variables that will guide the personalization process of behavioural interventions (De Bra, 2017). For instance, people’s privacy attitude has been often proposed as an adaptation means in the design of privacy wizards. Under this approach, users are classified into fundamentalists, pragmatists, or unconcerned, and their privacy policies adjusted to each of these categories (Knijnenburg, 2014; Alekh, 2018). As result, fundamentalists receive strong privacy policies, moderate settings are assigned to pragmatists, and weak ones to unconcerned users. In this case, the user model is said to be explicit since it is generated out infor-

mation that the users provide before starting to use the system (e.g. in an attitude questionnaire). However, the need for explicit user input can be diminished when implicit models are automatically generated from large data sets (De Bra, 2017). Under this approach, the model is automatically obtained out of information that emerges from the interaction between the user and the system. Particularly, information such as likes, clicks, and comments is aggregated into an implicit model that guides the adaptation of the system's interventions (Figure 1).

In general, the application of AI to nudging solutions offers the potential of boosting the effectiveness of behavioural interventions. However, such effectiveness comes with a list of drawbacks inherited from the underlying principles of AI technologies. Indeed, as personalization in nudges increases, concerns related to automated decision making and profiling quickly arise along with issues of transparency, fairness, and explainability (Susser, 2019; Brundage et al., 2018). Consequently, the user model and adaptation mechanism underlying these nudges should be scrutable in order to prevent inaccurate, unfair, biased, or discriminatory interventions. This would not only improve the system's accountability but would also give insights to the users on how their personal data is being used to promote changes in their behaviour. Over the last years, explainable AI (XAI) has shed light on many of these points and introduced methods for achieving transparent and accountable solutions. One example is the introduction of self-awareness mechanisms that endow deep learning systems with the capability to automatically evaluate their own beliefs and detect potential biases (Garigliano and Mich, 2019). However, the combination of AI with persuasion introduces additional challenges related to the impact that these technologies may have on the resulting behaviour of individuals and society in general (Müller, 2020; Susser, 2019). Hence, the definition of ethical guidelines, codes of conduct, and legal provisions are critical for guiding the development process of personalized nudges and for preventing a negative effect on people's well-being.

3.3 Ethical Guidelines

Although many have shown excitement about nudges and their applications, others consider this type of persuasive technologies as a potential threat to the users' agency, and autonomy (Renaud and Zimmermann, 2018; Susser, 2019). Particularly, some argue that nudges do not necessarily contribute to users' welfare and could even be used for questionable and unethical purposes. For instance, a mobile application can

nudge users to enable their phone's location services with the excuse of improving the experience within the app when, in fact, the main purpose is monitoring their movements. One case alike took place recently in China during the outbreak of the Coronavirus: the Chinese government implemented a system to monitor the virus's expansion and notify citizens in case they need to self-quarantine². Such a system generates a personal QR code which is scanned by the police and other authorities to determine whether someone is allowed into subways, malls, and other public spaces. However, although the system encourages people to provide personal information such as location and body temperature on behalf of public safety, experts suggest that this is another attempt by the Chinese government to increase mass surveillance.

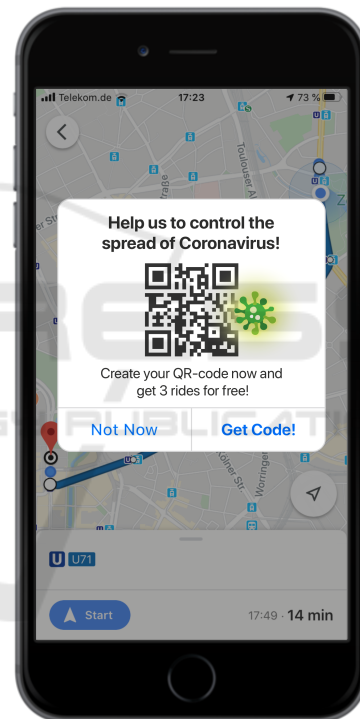


Figure 2: Nudge for monitoring the Corovavirus spread.

One distinctive aspect of nudges is that they are applied in a decision-making context. That is, they are used to encourage the selection of one alternative over others with the aim of maximizing people's welfare (Weinmann et al., 2016). In the case of the QR system introduced in China, its adoption was made mandatory by the government. Hence, it does not qualify as a nudge solution since this would normally

²Paul Mozur, Raymond Zhong and Aaron Krolik "In Coronavirus Fight, China Gives Citizens a Color Code, With Red Flags" *New York Times*, March 1, 2020. Accessed July 24, 2020.

involve the manipulation of the environment in which a decision is made while preserving users' autonomy and freedom of choice. However, one could imagine a nudge variant of such a system: the QR-code could be included as an optional feature inside a mobile application (e.g. public transport), and its use incentivized through a reward mechanism (e.g. a discount in the next trip) as illustrated in Figure 2. This and other instances of choice architectures give rise to ethical questions like "who should benefit from nudges?", "should users be informed of the presence of a nudge?" and "how nudges should (not) influence the users?". In line with this, (Renaud and Zimmermann, 2018) elaborated on a set of principles that privacy and security nudges should incorporate into their design to address these and other ethical concerns. Particularly, they introduced check-lists that designers can use to verify if their nudging solutions comply with principles such as justice, beneficence, and respect. For instance, to preserve users' autonomy, designers should ensure that all the original options are made available. This means that, if a nudge attempts to discourage people from installing privacy-invasive apps on their phones, then users should still have the option to install these apps if they wish to. Moreover, users should always be nudged towards behaviours that maximize their welfare rather than the interests of others. That is, choices that enclose a benefit for the designer (if any) should not be prioritized over those who do benefit the user.

4 RISK-BASED APPROACHES

As discussed, nudges can raise ethical concerns even if they are conceived for seemingly noble purposes like privacy and security. Furthermore, although personalization increases their effectiveness, it can also compromise users' privacy and autonomy. In this section, risk awareness is discussed and presented as a suitable means for developing appropriate nudging solutions. Particularly, it elaborates on how risk cues influence peoples' privacy decisions and how choice architectures may incorporate such cues into their design. Moreover, it discusses state-of-the-art solutions in which risk perception has been introduced as an adaptation variable for personalizing behavioural interventions.

4.1 The Role of Risk Cues

Risks are part of our daily life since there is always some uncertainty associated with the decisions we make. Moreover, it is our perception of risk which

often helps us to estimate the impact of our actions and influences our behaviour (Williams and Noyes, 2007). However, evaluating a large number of risk factors is often difficult due to the limited cognitive capacity of humans in general. Consequently, people often misjudge the consequences of their actions, behave unseemingly, and suffer unwanted incidents (Fischer, 2017). To avoid this, it is of utmost importance to increase individuals' sense of awareness and so their access to explicit and adequate risk information (Kim, 2017). This premise not only applies to decisions that are made in the real world but also in online contexts such as the disclosure of personal information. Particularly, self-disclosure is a practice which is usually performed under uncertainty conditions related to its potential benefits and privacy costs (Acquisti et al., 2015). However, average SNSs users find it difficult to perform proper estimations of the risks and benefits of their disclosures and, in turn, replace rational estimations with cognitive heuristics. For example, they often ground their privacy decisions on cues related to the platform's reputation (e.g. its size) or recognition (e.g. its market presence), among others (Marmion et al., 2017). All in all, the application of heuristics tends to simplify complex self-disclosure decisions. However, these heuristics can also undermine people's privacy-preserving behaviour since SNSs portray many trust-related cues, yet scarce risk information (Marmion et al., 2017). Furthermore, privacy policies are also devoid of risk cues which, in turn, hinder users' decisions related to consent on data processing activities (De and Imine, 2019). Consequently, even users with high privacy concerns may lack adequate means for conducting a rigorous uncertainty calculus.

4.2 Risk, Persuasion and AI

In general, risk awareness has a strong influence on people's behaviour and plays a key role in their privacy decisions. Therefore, the presence of risk cues is essential for supporting users in their self-disclosure practices. Under this premise, privacy scholars have introduced nudging solutions that aim to promote changes in people's privacy behaviour using risk information as a persuasive means. For example, (De and Le Métayer, 2018) introduced an approach based on attack-trees and empirical evidence to inform users of SNSs about the privacy risks of using lax privacy settings (e.g. the risks of having a public profile). Similarly, (Sanchez and Viejo, 2015) developed a method for automatically assessing the sensitivity degree of textual publications in SNSs (e.g. tweets or posts). Such a method takes into consideration the de-

gree of trust a user has in the targeted audience of a message (e.g. family, close friends, acquaintances) when determining its sensitiveness level. Furthermore, this approach is embedded in a system which notifies the users when privacy-critical content is being disclosed and suggests them to either restrict the publication's audience or remove/replace the sensitive terms with less detailed information. Nevertheless, the system ignores nuances in people's privacy goals and does not provide a mechanism for personalizing the interventions.

In order to guide the development of preventative nudges, (Díaz Ferreyra et al., 2018) introduced three design principles. The first one, *adaptivity*, refers to the importance of personalized interventions in creating engagement between the nudge and its users. Particularly, personalization is considered key for engaging individuals in a sustained learning process about good privacy practices. The second one, *viscerality*, highlights the importance of creating a strong and appreciable connection between users and their personal data. This principle is grounded on empirical evidence showing that, in general, users take conscience about the value of their data only after they suffer an unwanted incident (e.g. phishing or financial fraud). Finally, the principle of *sportiveness*, suggests that nudging solutions to cybersecurity should recommend countermeasures or coping strategies that users can put into practice to safeguard their privacy. When it comes to sportiveness, the authors suggest that many of the current privacy-enhancing technologies such as access-control lists and two-step verification would qualify as countermeasures and that the role of nudges is to motivate their adoption. On the other hand, they also suggest that adaptation and viscerality can be achieved by defining a user model which reflects individuals' risk perception (Díaz Ferreyra et al., 2020). Particularly, they introduced a user model consisting of a risk threshold which is updated as behavioural interventions are accepted or ignored by the end-user. By doing so, the nudge adapts to the individual privacy goals of the users and increases the effectiveness of its interventions. That approach was

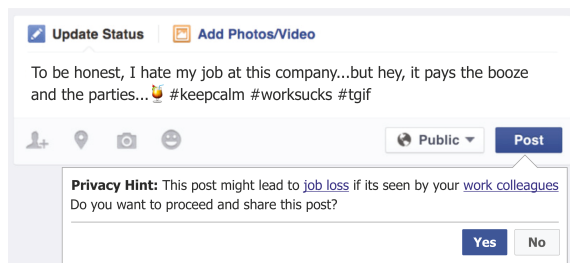


Figure 3: Personalized interventions for online-self disclosure (Díaz Ferreyra et al., 2020).

put into practice in the design of preventative technologies for SNSs as depicted in Figure 3. Particularly, this approach uses empirical evidence on regrettable self-disclosure experiences to elaborate risk patterns and shape behavioural interventions. Such risk-based interventions aim to encourage the use of friend lists for controlling the audience of textual publications.

5 DISCUSSION AND CONCLUSION

Overall, social interaction across SNSs demands making privacy decisions on a frequent basis. However, online self-disclosure, as well as the evolving and ongoing nature of privacy choices, seem to be out of the scope of privacy regulations. Instead, data protection frameworks tend to focus more on issues related to consent and overlook (sometimes deliberately) the importance of providing the necessary means for performing an adequate privacy calculus. Consequently, service providers limit themselves to the definition of instruments for obtaining consent (e.g. privacy policies) leaving the rational estimations of privacy risks to the individual discretion of the users. However, as discussed throughout this paper, such estimations are often impaired by cognitive and motivational biases which tend to outweigh anticipated benefits over potential risks. Hence, users are prone to experience regret after disclosing personal information in SNSs due to false estimations and optimistic biases. Furthermore, because of such spurious estimations, people might end up sharing their private information with untrusted audiences and increase their chances of suffering social engineering attacks. Thus, the incorporation of awareness mechanisms is of utmost importance for supporting users' self-disclosure decisions and mitigating the likelihood of unwanted incidents.

At their core, social engineering countermeasures require promoting behavioural changes among the users of SNSs. Hence, nudging techniques are of great value for the design of technical solutions that could guide individuals towards safer privacy decisions. Furthermore, AI-based approaches can improve the effectiveness of such solutions by endowing them with adaptation and personalization features that address the individual goals and concerns of the users. However, despite its promising effects, the combination of AI together with persuasive means can result in unethical technological designs. In principle, this is because nudges can mislead people towards a behaviour which is not necessarily beneficial for them. In this case, ethical guidelines are quite

clear since they stress the importance of designing choice architectures that are transparent and tend to maximize people's welfare. However, the question that still remains unclear is whether platforms should keep self-regulating the design of these technologies, or if public actors should harness the application of ethical standards in order to safeguard individuals' agency and autonomy. Either way, conducting an adequate *social welfare impact assessment* of persuasive AI technologies is crucial to determine their effects (positive or negative) on human behaviour at large.

Another controversial point is that users are not always aware of the presence of a nudge since persuasive means target primarily people's automatic and subconscious processing system. However, ethical approaches should allow people to explicitly recognize the presence of the nudge and the influence it is aiming to excerpt. Hence, choice architectures should, when possible, introduce mechanisms that target individuals' reflective reasoning in order to avoid potential manipulation effects. As discussed in Section 4.1, social engineering countermeasures can achieve this by incorporating risk information and cues in their design. However, even risk-related information can be subject to manipulation if not framed accordingly. That is, when high-risk events are portrayed as low-risk situations and vice-versa. Furthermore, biases can also be introduced if the likelihood and consequence levels of unwanted incidents are not properly estimated and quantified. Hence, guidelines for the correct estimation of risks together with ethical approaches for their communication should be further investigated, introduced, and guaranteed.

ACKNOWLEDGEMENTS

This work was partially supported the H2020 European Project No. 787034 "PDP4E: Privacy and Data Protection Methods for Engineering" and Canada's Natural Sciences and Engineering Research Council (NSERC).

REFERENCES

- Acquisti, A., Adjerid, I., Balebako, R., Brandimarte, L., Cranor, L. F., Komanduri, S., Leon, P. G., Sadeh, N., Schaub, F., Sleeper, M., Wang, Y., and Wilson, S. (2017). Nudges for Privacy and Security: Understanding and Assisting Users' Choices Online. *ACM Computing Surveys (CSUR)*, 50(3):44.
- Acquisti, A., Brandimarte, L., and Loewenstein, G. (2015). Privacy and human behavior in the age of information. *Science*, 347(6221):509–514.
- Aïmeur, E., Diaz Ferreyra, N. E., and Hage, H. (2019). Manipulation and Malicious Personalization: Exploring the Self-Disclosure Biases Exploited by Deceptive Attackers on Social Media. *Frontiers in Artificial Intelligence*, 2:26.
- Aïmeur, E., Hage, H., and Amri, S. (2018). The Scourge of Online Deception in Social Networks. In Arabia, H. R., Deligiannidis, L., Tinetti, F. G., and Tran, Q.-N., editors, *Proceedings of the 2018 Annual Conference on Computational Science and Computational Intelligence (CSCI' 18)*, pages 1266–1271. IEEE Computer Society.
- Albladi, S. and Weir, G. R. S. (2016). Vulnerability to Social Engineering in Social Networks: A Proposed User-Centric Framework. In *2016 IEEE International Conference on Cybercrime and Computer Forensic (ICCCF)*, pages 1–6. IEEE.
- Alekh, S. (2018). Human Aspects and Perception of Privacy in Relation to Personalization. *arXiv preprint arXiv:1805.08280*.
- Barev, T. J. and Janson, A. (2019). Towards an Integrative Understanding of Privacy Nudging - Systematic Review and Research Agenda. In *18th Annual Pre-ICIS Workshop on HCI Research in MIS (ICIS)*.
- boyd, d. (2010). *Social Network Sites as Networked Publics: Affordances, Dynamics, and Implications*, pages 47–66. Routledge.
- Briscoe, E. J., Appling, D. S., and Hayes, H. (2014). Cues to Deception in Social Media Communications. In *47th Annual Hawaii International Conference on System Sciences (HICSS)*, pages 1435–1443. IEEE.
- Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., Filar, B., et al. (2018). The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation. *arXiv preprint arXiv:1802.07228*.
- Bullée, J.-W. H., Montoya, L., Pieters, W., Junger, M., and Hartel, P. (2018). On the anatomy of social engineering attacks - A literature-based dissection of successful attacks. *Journal of Investigative Psychology and Offender Profiling*, 15(1):20–45.
- De, S. J. and Imine, A. (2019). On Consent in Online Social Networks: Privacy Impacts and Research Directions (Short Paper). In Zemhari, A., Mosbah, M., Cuppens-Bouahia, N., and Cuppens, F., editors, *Risks and Security of Internet and Systems*, pages 128–135. Springer International Publishing.
- De, S. J. and Le Métayer, D. (2018). Privacy Risk Analysis to Enable Informed Privacy Settings. In *2018 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, pages 95–102. IEEE.
- De Bra, P. (2017). Challenges in User Modeling and Personalization. *IEEE Intelligent Systems*, 32(5):76–80.
- Díaz Ferreyra, N., Meis, R., and Heisel, M. (2018). At Your Own Risk: Shaping Privacy Heuristics for Online Self-disclosure. In *2018 16th Annual Conference on Privacy, Security and Trust (PST)*, pages 1–10.
- Díaz Ferreyra, N. E., Kroll, T., Aïmeur, E., Stieglitz, S., and Heisel, M. (2020). Preventative Nudges: Introducing

- Risk Cues for Supporting Online Self-Disclosure Decisions. *Information*, 11(8):399.
- Egelman, S., Sotirakopoulos, A., Muslukhov, I., Beznosov, K., and Herley, C. (2013). Does my password go up to eleven? The impact of password meters on password selection. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2379–2388.
- Fischer, A. R. H. (2017). Perception of Product Risks. In Emilien, G., Weitkunat, R., and Lüdicke, F., editors, *Consumer Perception of Product Risks and Benefits*, pages 175–190. Springer International Publishing, Cham.
- Garigliano, R. and Mich, L. (2019). Looking Inside the Black Box: Core Semantics Towards Accountability of Artificial Intelligence. In *From Software Engineering to Formal Methods and Tools, and Back*, pages 250–266. Springer.
- Hage, H., Aïmeur, E., and Guedidi, A. (2020). Understanding the landscape of online deception. In *Navigating Fake News, Alternative Facts, and Misinformation in a Post-Truth World*, pages 290–317. IGI Global.
- Hamari, J. and Koivisto, J. (2013). Social Motivations To Use Gamification: An Empirical Study Of Gamifying Exercise. In *Proceedings of the 21st European Conference on Information Systems (ECIS)*.
- Howah, K. and Chugh, R. (2019). Do We Trust the Internet?: Ignorance and Overconfidence in Downloading and Installing Potentially Spyware-Infected Software. *Journal of Global Information Management (JGIM)*, 27(3):87–100.
- Kim, H. K. (2017). Risk Communication. In Emilien, G., Weitkunat, R., and Lüdicke, F., editors, *Consumer Perception of Product Risks and Benefits*, pages 125–149. Springer International Publishing, Cham.
- Knijnenburg, B. P. (2014). Information Disclosure Profiles for Segmentation and Recommendation. In *Symposium on Usable Privacy and Security (SOUPS)*.
- Krämer, N. C. and Schäwel, J. (2020). Mastering the challenge of balancing self-disclosure and privacy in social media. *Current Opinion in Psychology*, 31.
- Krombholz, K., Hobel, H., Huber, M., and Weippl, E. (2015). Advanced social engineering attacks. *Journal of Information Security and applications*, 22:113–122.
- Marmion, V., Bishop, F., Millard, D. E., and Stevenage, S. V. (2017). The Cognitive Heuristics Behind Disclosure Decisions. In *International Conference on Social Informatics*, pages 591–607. Springer.
- Masur, P. K. (2019). Privacy and self-disclosure in the age of information. In *Situational Privacy and Self-Disclosure*, pages 105–129. Springer.
- Müller, V. C. (2020). Ethics of Artificial Intelligence and Robotics. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, fall 2020 edition.
- Peer, E., Egelman, S., Harbach, M., Malkin, N., Mathur, A., and Frik, A. (2020). Nudge Me Right: Personalizing Online Nudges to People’s Decision-Making Styles. *Computers in Human Behavior*, 109:106347.
- Penni, J. (2017). The Future of Online Social Networks (OSN): A Measurement Analysis Using Social Media Tools and Application. *Telematics and Informatics*, 34(5):498–517.
- Petkos, G., Papadopoulos, S., and Kompatsiaris, Y. (2015). PScore: A Framework for Enhancing Privacy Awareness in Online Social Networks. In *2015 10th International Conference on Availability, Reliability and Security*, pages 592–600. IEEE.
- Renaud, K. and Zimmermann, V. (2018). Ethical guidelines for nudging in information security & privacy. *International Journal of Human-Computer Studies*, 120:22–35.
- Sanchez, D. and Viejo, A. (2015). Privacy Risk Assessment of Textual Publications in Social Networks. In *Proceedings of the International Conference on Agents and Artificial Intelligence - Volume 1, ICAART 2015*, pages 236–241, Setubal, PRT. SCITEPRESS - Science and Technology Publications, Lda.
- Susser, D. (2019). Invisible Influence: Artificial Intelligence and the Ethics of Adaptive Choice Architectures. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 403–408.
- Tsikerdekis, M. and Zeadally, S. (2014). Online Deception in Social Media. *Communications of the ACM*, 57(9):72.
- Turland, J., Coventry, L., Jeske, D., Briggs, P., and van Moorsel, A. (2015). Nudging towards Security: Developing an Application for Wireless Network Selection for Android Phones. In *Proceedings of the 2015 British HCI Conference*, British HCI ’15, pages 193–201, New York, NY, USA. Association for Computing Machinery.
- Vassileva, J. (2012). Motivating participation in social computing applications: a user modeling perspective. *User Modeling and User-Adapted Interaction*, 22(1-2):177–201.
- Wang, Y., Leon, P. G., Scott, K., Chen, X., Acquisti, A., and Cranor, L. F. (2013). Privacy Nudges for Social Media: An Exploratory Facebook Study. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 763–770.
- Wang, Y., Norcie, G., Komanduri, S., Acquisti, A., Leon, P. G., and Cranor, L. F. (2011). “I regretted the minute I pressed share”: A Qualitative Study of Regrets on Facebook. In *Proceedings of the 7th Symposium on Usable Privacy and Security, SOUPS 2011*, pages 1–16. ACM.
- Warberg, L., Acquisti, A., and Sicker, D. (2019). Can Privacy Nudges be Tailored to Individuals’ Decision Making and Personality Traits? In *Proceedings of the 18th ACM Workshop on Privacy in the Electronic Society, WPES’19*, pages 175–197. ACM.
- Weinmann, M., Schneider, C., and vom Brocke, J. (2016). Digital Nudging. *Business & Information Systems Engineering*, 58(6):433–436.
- Williams, D. J. and Noyes, J. M. (2007). How does our perception of risk influence decision-making? Implications for the design of risk information. *Theoretical Issues in Ergonomics Science*, 8(1):1–35.