Modelling Responsible Digital Security Behaviour for Countering Social Media Manipulation

Clara Maathuis[1] and Sabarathinam Chockalingam[2]
[1]Open University of the Netherlands, Heerlen, Netherlands
[2]Institute for Energy Technology, Halden, Norway
clara.maathuis@ou.nl
sabarathinam.chockalingam@ife.no

Abstract: While the digital environment, and in particular social media, surrounds not only human's identity and its societal functions projection, e.g., institutional and financial aspects, it also captures both individual and collective thoughts regarding former, ongoing, and future concepts, trends, and incidents placed in the physical world, in the digital environment, or in both which could impact both individual and collective consciousness, behaviour, and attitude towards different dimensions of reality. Accordingly, an initial attempt to define and model responsible digital security behaviour was made and ongoing discourses and AI-based solutions for tackling and containing social manipulation mechanisms exist in this domain. Noteworthily is that dedicated attention to understanding and modelling responsible digital security behaviour in social media for tackling and/or countering social media manipulation, e.g., disinformation and misinformation, still lacks. To this end, this research aims (i) to capture the factors influencing user behaviour towards tackling and/or countering social media manipulation, (ii) to build a Machine Learning model that assesses user's responsibility in relation to tackling and/or countering social media manipulation mechanisms, and (iii) propose a set of socio-technical recommendations for building resilience to such mechanisms. To accomplish these research objectives, a Design Science Research methodological approach is taken by designing, developing, and evaluating the model proposed through exemplification. Finally, this research aims to enhance digital security awareness and resilience to social media manipulation of users and policy decision-makers to manage and further extend in a responsible and safe way the digital environment.

Keywords: information operations, cyber operations, cyber security, digital security, responsible security, digital security behaviour, Bayesian networks.

1. Introduction

With each day more people and organizations are projecting their identity, vision, and (diverse dimensions of) their existence in the social media realm. On the one hand, this further facilitates the grow of information together with exchange and communication among different users, communities, and societies (Cinelli et al. 2019). On the other hand, focusing on the scale and intensity of its impact, the information communicated and exchanged can be either real or false through intentional alteration and manufacturing or unintentional through, e.g., mistake. This could lead to implications and consequences like discreditation, confusion, behaviour change and negatively influencing decision-making processes, time, and resources allocated during crisis, and alter public perception in social, political, economic, and health settings (Bontcheva et al. 2020; Chockalingam & Maathuis, 2022). Such social manipulation campaigns are the ones conducted for addressing and influencing people's perspectives on historical and political aspects like in the wars in Syria and Ukraine by specialized cyber troops (Filton, 2016; Bradshaw & Howard, 2018; Pavlik, 2022), on climate change issues (Mavrodieva et al. 2019), and on symptom, vaccine composition, and vaccination impact during the Covid-19 pandemic and Infodemic (Li et al. 2022). While such issues are continued to be addressed by researcher and practitioner communities together with dedicated social platforms, organizations, and governments through dedicated solutions like campaigns, alerts, fact checking, and filtering algorithms, it all comes down to the level of user behaviour. The user must be aware and proactively address social manipulation by building resilience and further effectively tackling and/or countering it. This represents an important knowledge gap that this research intends to tackle which may lead to a new research path in this direction. Accordingly, this research aims to (i) gather and cluster influential factors that correspond to the responsible digital security behaviour in social media, and (ii) develop an AI-based model for assessing such behaviour. Hence, the following research questions are defined:

- What are the influential factors representing the responsible digital security behaviour when tackling and/or countering social media manipulation?
- How to develop an AI assessment model that classifies the responsible digital security behaviour in social media as being responsible or not?

Hence, while through the first research question features of the responsible digital security behaviour are captured, they are further used in the second research question to build an AI-model that assesses this behaviour. To tackle these questions, a Design Science Research methodological approach is considered based

on a literature review conducted in the digital/cyber security, social media manipulation, psychology, behaviour and cognitive science, and AI domains. Hence, the contributions of this research can be summarized as follows:

- Identifies and categorizes the factors contributing to building a responsible digital security behaviour in combating social media manipulation.
- Addresses the need for building an AI model for producing and enhancing awareness and resilience to social media manipulation.

The reminder of this article is structured as follows. Section 2 settles the context while discussing relevant scientific studies and governmental strategies and policies. Section 3 highlights the research methodology used for achieving the aim of this research. Section 4 presents the working definition of this research and prepares the terrain for Section 5 where influencing factors are gathered and clustered into three different categories. Section 6 discusses the developed model and evaluation through demonstration. Conclusively, Section 7 summarizes the findings of this research and presents future research ideas.

2. Social Media Manipulation Related Research

Taking into consideration the complex nature and the multi-facetted aspects of this research, multidisciplinary research is conducted capturing relevant characteristics from the digital/cyber security, social media manipulation, psychology, behaviour and cognitive science, and AI domains. Furthermore, this research is positioned in the existing body of knowledge.

The first step in assessing and further building user and/or social platform capacity for countering social media manipulation is awareness. Accordingly, Ki-Aries & Faily (2017) identify human factors for building persona-centered information security awareness stressing that such a perspective allows building a security capacity (skills) adaptive in time for generating a positive contribution when avoiding, containing, or reducing risks and corresponding malicious incidents. Next to awareness, behavioural change to existing or possible future risks is necessary. On this, Perski et al. (2017) conduct a systematic review on digital behaviour change and engagement. They point out among others, the importance of psychological influencing factors like experience, motivation, and self-efficacy. Bada et al. (2019) stress the need for including human factors, and particularly, psychological factors when building cyber security awareness campaigns and solutions. Moreover, Mashiane & Kritzinger (2018) propose a conceptual taxonomy for cyber security behaviour to categorize it including security awareness and possible naïve mistakes that users could make. Kamarulzaman, Shuhidan & Wahid (2022) propose a framework for capturing the effect of information security behaviour on the relation between digital literacy and information security culture including factors, e.g., critical thinking and problems solving, which are relevant to this research.

Going further into the social media realm, Ansar & Goswami (2021) conduct a survey on characteristics and detection methods applied to fake news and consider that their spread is done either through malicious accounts or through echo chambers. Such detection mechanisms rely on evaluating the content, source, and social response to it. Particularly, Caled & Silva (2022) conceptualize social media manipulation mechanisms such as disinformation and misinformation, further reflecting on cognitive vulnerabilities like bias that influence human behaviour and address the necessity of developing solutions based on multidisciplinary research. Traberg, Roozenbeek & van der Linden (2022) bring back attention to an old socio-psychological theory, i.e., the inoculation theory, for immunizing governments and social media companies by pre-exposing them with different social manipulation doses. It means building relevant scenarios from where they can learn how to recognize and combat social media manipulation. Regarding social media manipulation, Pennycook & Rand (2021) investigate the reasons why people believe and further share false or highly misleading information/news online. They emphasize that people are better judging truth from falsehood when evaluating political concordant news, while poor judgement is related to their lack of careful reasoning and when using heuristics like familiarity. At the same time, they also acknowledge that there is a substantial disconnection between what people believe and what they share in social media due to inattention, e.g., not properly checking the source of information. Effron & Helgason (2022) discuss three psychological factors that contribute to countering social media manipulation: partisanship, imagination, and repetition, and directly relate them to current social constructs, i.e., political polarization, alterative facts being pushed by leaders, and technology that amplifies social media manipulation. Furthermore, Reisach (2021) assesses the responsibility that social media platforms have in regards with countering disinformation and societal polarization. They also scrutinize that social media platforms have the capacity to minimize societal risks by taking responsible action in corresponding societal

domains, i.e., human rights, education, and transparency of algorithmic decisions. The role of social media platforms on countering manipulation mechanisms like disinformation is analyzed by Horowitz et al. (2022). Herein, the authors stress the importance and responsibility that social platforms have for producing fact-based content and building intelligent anti-disinformation solutions. Specifically on social media content, Schreiner, M., Fischer, T., & Riedl (2021) identify its role and relevant characteristics together with its relation to user's emotions and corresponding emotional responses. Moreover, Moran, Muzellec & Johnson (2019) analyze user engagement in social media and under this umbrella they capture in a conceptual model the following actions taken on content: clicks, links, comments, and shares. At the same time, several AI-based solutions continue to be developed and proposed in different settings as discussed in the comprehensive review conducted on deep learning detection solutions by (Mridha et al. 2021). To this end, Maathuis & Chockalingam (2022a) proposed a definition and an AI-based assessment model for responsible digital security behaviour which clusters the influential factors on such behaviour as human, infrastructure, and policy-based. The concept introduced, and the implemented model represent the basis of this research on top of which different discussed studies contribute to.

At a governance level, ITU (International Telecommunication Union) released a report on countering digital disinformation while the freedom of expression is respected by analyzing different international disinformation responding mechanisms together with the impact on the right of freedom of opinion and expression, touching as well corresponding concepts like online privacy and dignity issues Bontcheva et al. (2020). The European Commission tackles disinformation and misinformation through different policies, strategies, and programs for protecting values and democratic systems, e.g., the Strengthened Code of Practice on Disinformation, the Action plan on disinformation, and the European Democracy Action Plan were developed (EC, 2022). NATO continues to maintain its readiness to actively deal with disinformation and misinformation campaigns and operations "in a unified manner". This is done through fact-based, credible public communications, debunking false narratives, and further sharing proper and transparent "information, insight, and best practice" (NATO, 2020). At the same time, to mention some efforts of social media platforms, e.g., the identification and removal of accounts conducting inauthentic engagement on Facebook, Instagram, TikTok, and YouTube platforms. However, further efforts on transparency, accountability, and actively countering social manipulation are necessary (Bay et al. 2020). These efforts reflect the existing necessity of understanding and strengthening user's behaviour by building corresponding social media manipulation combating mechanisms and solutions.

3. Research Methodology

To capture and cluster the influential factors that contribute to developing responsible digital security behaviour for countering social media manipulation, a multidisciplinary stance is adopted. Taking into consideration the fact that the aim of this research reasons with a deep societal need and its underlying technicality, this research follows the Design Science Research methodology (Peffers et al. 2007; Peffers, Tuunanen & Niehaves, 2018). In terms of the model development, we adapt and extend the Responsible Digital Security Behaviour model proposed by Maathuis & Chockalingam (2022a) to combat social media manipulation. Consequently, the model proposed represents to the best of our knowledge the first modelling effort in this direction while stressing the necessity of further building explainable and responsible AI-based models for tackling societal challenges (Maathuis, 2022a; Maathuis, 2022b). Consequently, the following research activities are taken:

Problem Identification and Solution Objective**:** In recent years, the number, complexity, scale, and impact of social media manipulation has significantly increased reaching from precisely smaller target audiences to broader dynamic and imprecise (target) audiences reflecting direct social and technological engagement advancements (Maathuis, Pieters & van den Berg, 2018b). In this area, next to stakeholders like social media platforms, dedicated organizations, and governments, users must be digitally aware and responsible for properly preventing and countering social media manipulations mechanisms. Accordingly, through a literature review in the domains involved conducted on scientific databases like IEEE, ACM, and Springer using combinations of keywords like 'social manipulation', 'disinformation', 'misinformation', 'digital awareness', 'security behaviour', and 'AI', this research aims to capture influential factors for developing a model for representing digital security behaviour in social media for combating social media manipulation.

Design and Development: The factors are identified and clustered, and a Bayesian Belief Network model is designed and developed in this research considering the model introduced by Maathuis & Chockalingam (2022a). That model considers three clusters of factors, namely *human, infrastructure* and *policy* while tackling general digital security aspects. In the model introduced in this research, all categories of factors are extended while new ones particular to the social media domain are added. Specifically, the only common cluster between

the initial model and the one proposed in this research is the *human factors* model, but the one proposed here add other factors like *Social_influence* and *Cognitive_bias*, while the other two clusters are different: *infrastructure* and *policy clusters* from the initial model are replaced with *content* and *social media clusters* in the model proposed in this article. The choice for building the model proposed in this research using BBN is their nature and capability of representing knowledge characterizing a domain or a problem, representing probabilistic relationships between characterizing variables (factors in our case), reasoning about uncertain information and aspects, i.e., social media activities, and predictive modelling, i.e., being able to predict in which situation one's behaviour could be considered digitally responsible in social media. To add to these benefits of building BBN models are their easy of use in different decision-making processes, their explainability power since they are easy to understand and interpret, and their transparency characteristic since they support verification and validation using different types of data and scenarios generated by both real incidents and experts.

Demonstration and Evaluation**:** The model proposed is demonstrated and evaluated through demonstration on two different instantiations that capture two different user settings, i.e., the first one which reflects a responsible behaviour versus the second one which reflects an irresponsible behaviour towards countering social media manipulation.

Communication**:** The factors, model, and results obtained are disseminated through this article and corresponding presentations.

4. Responsible Digital Security Behaviour Countering Social Media Manipulation

Grounded on disintermediation and technological advancements, efforts built for influencing, disturbing, or damaging decision-making processes, opinion, belief, and/or behaviour of a target audience imply either a singular or a combination of more strategies and mechanisms like disinformation and misinformation (Cinelli et al. 2019). Such mechanisms should be flexible and tailorable and are labelled in the context of social media as social media manipulation. A few such examples are generating confusion, tensions, and conflict among the components of target society, and discrediting, destabilizing, undermining, or promoting specific (sub)groups of target society (Mazarr et al. 2019). Since social media manipulation increased in scale, volume, and distribution speed (Kapantai et al. 2021), both increased conventional media exploitation and exposure is created and given to, respectively (Fitzpatrick, 2018; Bada et al. 2019). However, confusion about the meaning of specific social media manipulation mechanisms like disinformation and misinformation still exists. Concretely, these mechanisms are macro-concepts as they include specific techniques like deep fakes, fake news, and rumors. Accordingly, disinformation refers to the intentional use of altered or fabricated information while misinformation implies the use of misleading or inaccurate information in an unintentional way (Caled & Silva, 2022; Guess & Lyons, 2020; Petratos, 2021). To tackle these mechanisms, apart from the detection and countering solutions developed by researchers plus social media platforms and organizations, awareness, and resilience of global users through instruments like inoculation theory and prebunking for pre-emptively building cognitive, social, psychological, and emotional resistance to such threats, are necessary (Traberg, Roozenbeek & van der Lindern, 2022; Fard & Maathuis, 2021). Such instruments call for enhancing existing or building new digital security behaviour in a responsible way in relation to tackling and/or countering social manipulation mechanisms. To this end, this research extends the definition for responsible digital security behaviour which points to how the agent acts in relation to a digital system in a given (digital) context respecting agreed norms, procedures, and values of agent's organization Maathuis & Chockalingam (2022a). This implies that when referring to responsibility we mean in this research the combination of ethical, legal, and social norms and values that both organizations and agents have. Hence, the working definition for responsible digital security behaviour in social media is considered:

Responsible Digital Security Behaviour in relation to tackling and/or countering social media manipulation is the way an agent (e.g., user) acts in relation to a social media component (e.g., content, other accounts, policy) in each (digital) context with respect to the agreed norms, procedures, and values of the social media platform.

5. Clustering Influencing Factors

The previous section introduced the working definition of this research. Furthermore, factors that influence it are captured and clustered considering human, content, and social media dimensions together with the interaction design as well as goal, role-based persona design, and content aspects and model introduced by (Nielsen, 2015; Wu et al. 2019; Maathuis & Chockalingam, 2022a) and captured in Table 1 and further explained:

- Input variables: human factors which include user related aspects, content factors which contain relevant content characteristics, and social media factors which include tackling and/or countering mechanisms in a socio-technical perspective.
- Output variable (target): Influence which can be responsible or irresponsible.

Table 1. Influential factors

Human Factors (Perski et al., 2017; Ki-Aries & Faily, 2017; Mashiane & Kritzinger 2018, 2018; Chowdhury, Adam & Skinner, 2019; Bontcheva et al., 2020; Pennycook & Rand, 2021; Effron & Helgason, 2022; Kamarulzaman, Shuhidan & Wahid, 2022; Caled & Silva, 2022)	• Experience (High, Medium, Low) • Critical thinking sand problem solving skills (High, Medium, Low) • Belief and motivation (High, Medium, Low) • Security awareness (True, False) • (Naïve) mistakes (True, False) • Self-efficacity/Response efficacity (High, Medium, Low) • Time/space pressure (True, False) • Social influence/Peer behaviour (High, Medium, Low) • Emotional conditions (True, False) • Cognitive bias (High, Medium, Low) • Morality (High, Medium, Low)
Content Factors (Wu et al., 2019; Moran, Muzellec & Johnson, 2019; Reisach, 2021; Ansar & Goswami, 2021; Schreiner, Fischer & Riedl, 2021; Traberg, Roozenbeek & van der Linden 2022; Horowitz et al., 2022)	• Source check (True, False) • Information evaluation (True, False) • Content engagement/Social response (High, Medium, Low)
Social Platform Factors (Reisach, 2021; Effron & Helgason, 2022; Horowitz et al., 2022; Caled & Silva, 2022)	• Policy application/action (True, False) • Risk perception in relation to sharing/content amplification (High, Medium, Low) • Countering/Combating mechanism effectiveness (High, Medium, Low) • Countering/Combating mechanism management (High, Medium, Low)

6. Model Design and Evaluation

Once the factors were identified and clustered, the modelling process starts by first briefly discussing the AI technique used, i.e., Bayesian Belief Networks. This represents a powerful formalism for automated reasoning with uncertainty incorporating probabilistic information and facilitating decision making processes in a diverse plethora of domains and applications (van der Gaag, 1996). This technique is used for building the model proposed in this research as it allows the incorporation of both qualitative and quantitative approaches in a complex and uncertain environment like social media. Accordingly, the qualitative component is a Directed Acyclic Graph which contains variables and relations/edges between them, and the quantitative component relates to the CPT (Conditional Probability Tables) which captures conditional probabilities for all child-parent variables states (Marcot & Penman, 2019), as depicted in Figure 1. As this research predicts the influence of factors identified, then it embeds a predictive reasoning approach, i.e., reasoning from cause (upper layer) to effect (lower layer).

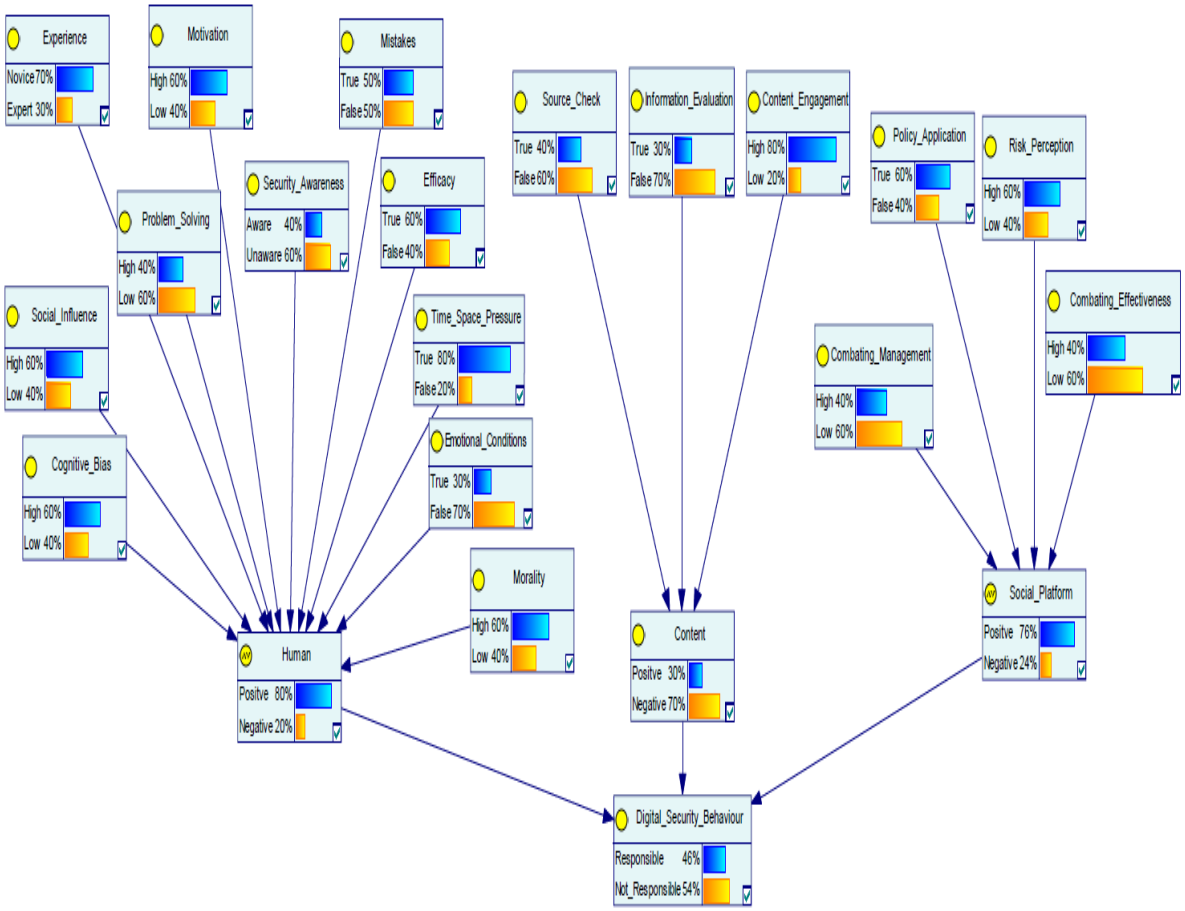

Figure 1. Proposed model without evidence provided.

In this model example values are used to demonstrate its functional mechanism while further evidence is provided on concrete cases of responsible versus irresponsible behaviour captured in Figure 2 and Figure 3.

In the first case shown in Figure 2, evidence is set in the upper layer as follows: Experience = 'Low', Security awareness = 'False', Emotional conditions = 'True', Cognitive bias = 'High', Information evaluation = 'False', and Risk perception = 'Medium'. Hence, the posterior probability (57%) is computed for the Influence target variable to 'Not_responsible'. This implies that the user shows an irresponsible digital security behaviour in relation to combating social media manipulation as it can be influenced by emotional narratives, staged videos, and polluted data (Bontcheva et al. (2020). Consequently, the user should be trained either based on individual initiative or through a collective initiative taken by social media, dedicated organizations, and governments, to be aware, alert, and effective in such situations considering a diverse plethora of possible situations in different manipulation contexts and settings (Horowitz et al. 2022), e.g., concrete, and transparent incident analysis and implications based on the dynamics involved in the ongoing war in Ukraine.

In the second case depicted in Figure 3, evidence is set in such a way: Experience = 'High', Critical thinking = 'High', Self-efficacy = 'High', Cognitive bias = 'Low', Information evaluation = 'True', and Risk perception = 'High'. Hence, the posterior probability (53%) is computed for the Influence target variable to 'Responsible'. This implies that the user shows a responsible digital security behaviour in relation to combating social media manipulation mechanisms since is difficult to be manipulated. Accordingly, the user should remain further alert and effective while remaining open to media governance initiatives like policies and advises to further maintain and further strengthen this behaviour (Reisach, 2021).

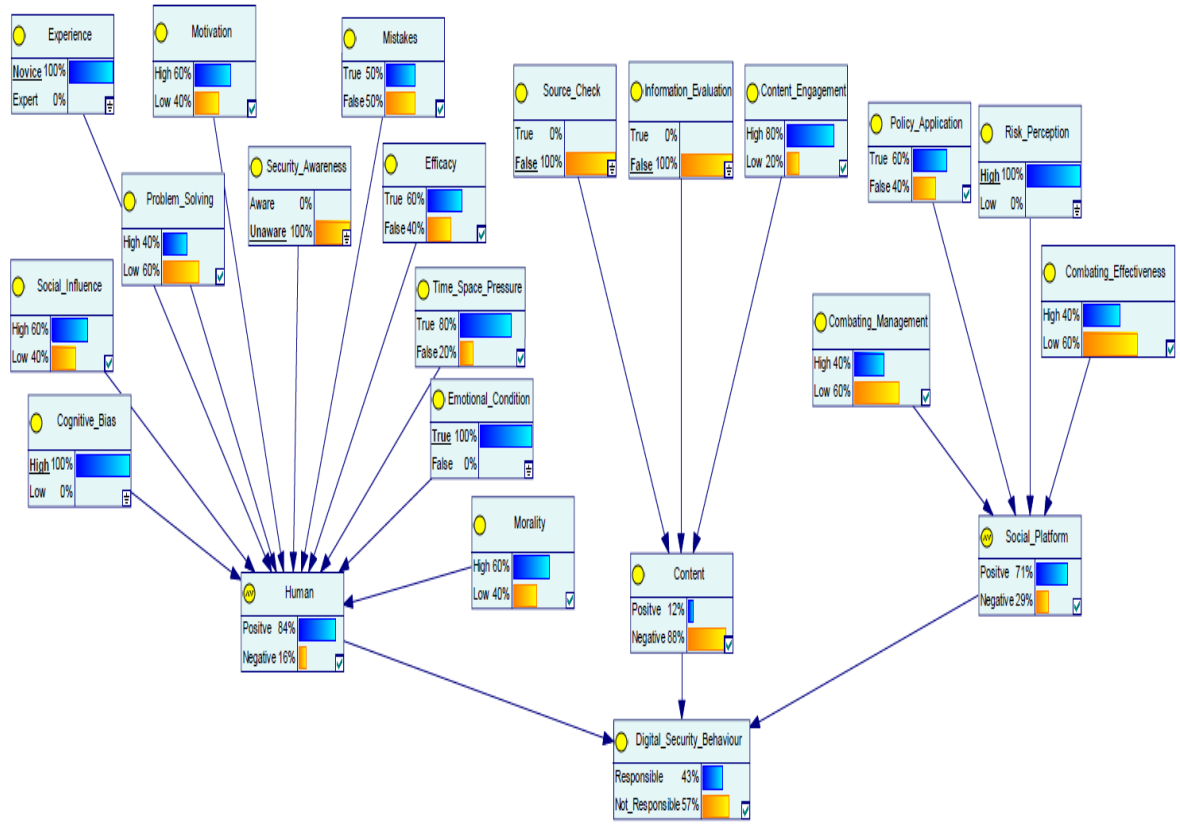

Figure 2. First demonstration

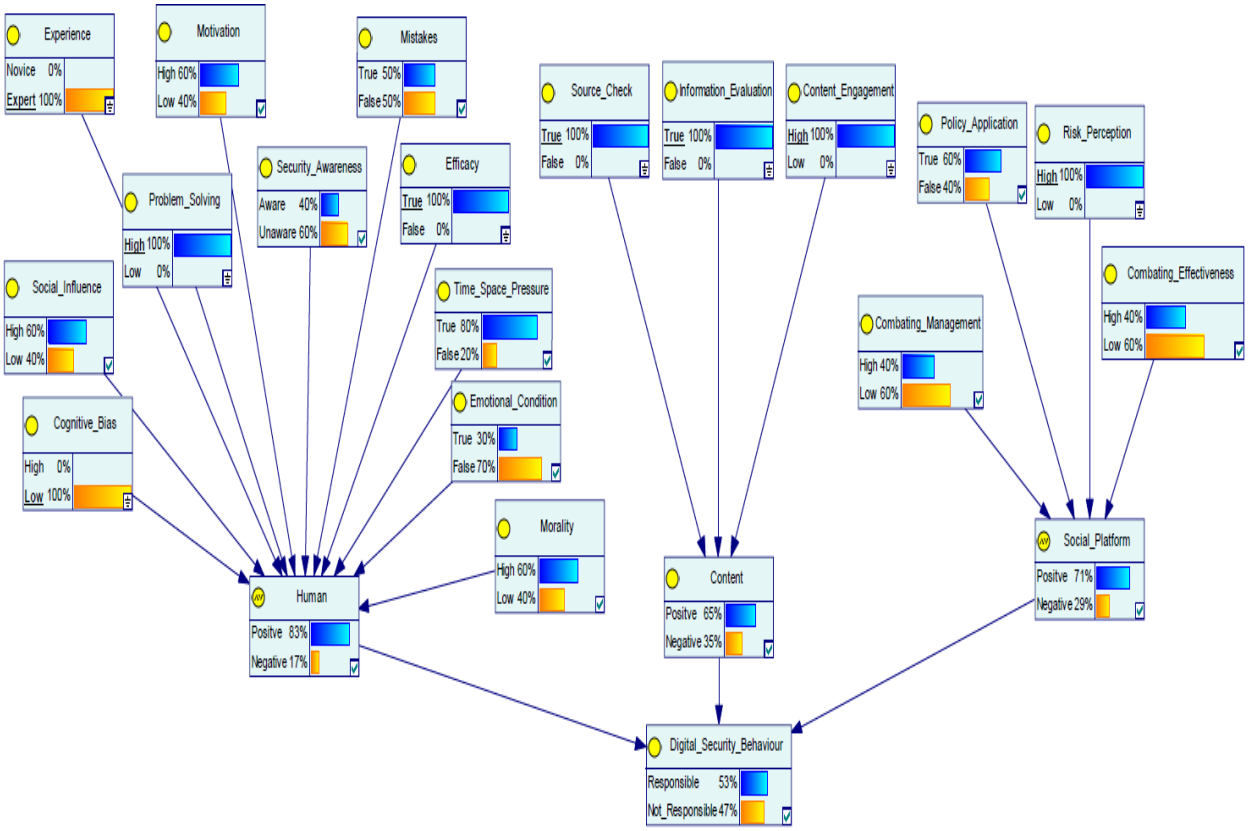

Figure 3. Second demonstration

7. Conclusions and Future Work

Seeing the progress, blend, and pervasiveness of technology (Jongepier & Klenk, 2022) in all societal domains allows one to understand that with and through information capturing evil or malicious intentional goals, manipulation could be inflicted in one's belief and behaviour in diverse digital environments like social media. This directly reflects exploiting users' vulnerabilities which could produce major implications and consequences to both targeted user audiences as well as to other collateral ones (Maathuis, Pieters & van den Berg, 2018a). To be aware and further build (intelligent) tackling and/or countering social media mechanisms, the four cornerstones of an engaging security culture need to be considered: responsibility, trust, communication, and cooperation (Ki-Aries & Faily, 2017). As the last two aspects represent a different type of research than the one conducted herein, the first two aspects are correlated and directly tackled in this research by building a responsible digital security behaviour model for social media using AI. Accordingly, multidisciplinary research is conducted using the Design Science Research methodology, and relevant influencing factors on this behaviour are gathered and clustered in three categories of factors: human, content, and social media platform. Moreover, the model proposed classifies the predicted behaviour as being responsible and irresponsible and is evaluated through demonstration based on synthetical data. The model can represent a basis for establishing proper user socio-technical recommendations for either building or enhancing the responsible digital security behaviour in relation to social media manipulation.

Furthermore, other dimensions and influence factors will be considered in future research through a socio-technical approach to capture, assess, and instantiate, e.g., governance, legal, and ethical factors to building and/or enhancing responsible digital security behaviour when dealing or countering social media manipulation. Conclusively, this research contributes to building responsible, transparent, and accountable AI-based solutions for developing awareness and resilience actively preventing, identifying, dealing, and countering social media manipulation which continues to increase in complexity, intelligence, and adaptivity.

References

Ansar, W., & Goswami, S. (2021). Combating the menace: A survey on characterization and detection of fake news from a data science perspective. *International Journal of Information Management Data Insights*, *1*(2), 100052.

Bada, M., Sasse, A. M., & Nurse, J. R. (2019). Cyber security awareness campaigns: Why do they fail to change behaviour? *arXiv preprint arXiv:1901.02672*.

Bay, S., Fredheim, R., Haiduchyk, T. & Dek, A. (2021). Social media manipulation 2021/2022: assessing the ability of social media companies to combat platform manipulation. NATO STRATCOM COE.

Bontcheva, K., Posetti, J., Teyssou, D., Meyer, T., Gregory, S., Hanot, C., & Maynard, D. (2020). Balancing act: Countering digital disinformation while respecting freedom of expression. *Geneva, Switzerland: United Nations Educational, Scientific and Cultural Organization*.

Bradshaw, S., & Howard, P. N. (2018). Challenging truth and trust: A global inventory of organized social media manipulation. *The computational propaganda project*, *1*, pp. 1-26.

Brangetto, P., & Veenendaal, M. A. (2016). Influence cyber operations: The use of cyberattacks in support of influence operations. In *2016 8th International Conference on Cyber Conflict (CyCon)* IEEE, pp. 113-126.

Caled, D., & Silva, M. J. (2022). Digital media and misinformation: An outlook on multidisciplinary strategies against manipulation. *Journal of Computational Social Science*, vol. *5*(1), pp. 123-159.

Chockalingam, S., & Maathuis, C. (2022). An Ontology for Effective Security Incident Management. In *International Conference on Cyber Warfare and Security* (Vol. 17, No. 1, pp. 26-35).

Cinelli, M., Conti, M., Finos, L., Grisolia, F., Novak, P. K., Peruzzi, A., ... & Quattrociocchi, W. (2019). (Mis) Information Operations. *Journal of Information Warfare*, vol. *18*(3), pp. 83-98.

Chowdhury, N. H., Adam, M. T., & Skinner, G. (2019). The impact of time pressure on cybersecurity behaviour: a systematic literature review. *Behaviour & Information Technology*, vol. *38*(12), pp. 1290-1308.

European Commission (2022). Tackling online disinformation. https://digital-strategy.ec.europa.eu/en/policies/online-disinformation.

Effron, D. A., & Helgason, B. A. (2022). The Moral Psychology of Misinformation: Why We Excuse Dishonesty in a Post-Truth World. *Current Opinion in Psychology*, 101375.

Fard, A. E., & Maathuis, C. (2021). Toward Capturing the Underlying Offensive Mechanisms of Social Manipulation: A Data Model Approach.

Fitton, O. (2016). Cyber operations and gray zones: Challenges for NATO. *Connections*, vol. *15*(2), pp. 109-119.

Fitzpatrick, N. (2018). Media manipulation 2.0: the impact of social media on news, competition, and accuracy.

Gill, T. D., & Ziolkowski, K. (2013). Non-intervention in the Cyber Context. *Peacetime Regime for State Activities in Cyberspace International Law. International Relations and Diplomacy. NATO CCDCOE, Tallinn*, pp. 217-238.

Guess, A. M., & Lyons, B. A. (2020). Misinformation, disinformation, and online propaganda. *Social media and democracy: The state of the field, prospects for reform*.

Horowitz, M., Cushion, S., Dragomir, M., Gutiérrez Manjón, S., & Pantti, M. (2022). A framework for assessing the role of public service media organizations in countering disinformation. *Digital Journalism*, vol. *10*(5), pp. 843-865.

Jongepier, F., & Klenk, M. (2022). The Philosophy of Online Manipulation. Routledge.

Kamarulzaman, M. S., Shuhidan, S. M., & Wahid, K. A. (2022, September). The Moderating Role of Information Security Behaviour (ISB) on the Relationship between Digital Literacy (DL) and Information Security Culture (ISC): A Proposed Research Framework. In *Proceedings* (Vol. 82, No. 1, p. 35). MDPI.

Kapantai, E., Christopoulou, A., Berberidis, C., & Peristeras, V. (2021). A systematic literature review on disinformation: Toward a unified taxonomical framework. *New media & society*, vol. *23*(5), pp. 1301-1326.

Ki-Aries, D., & Faily, S. (2017). Persona-centred information security awareness. *computers & security*, vol. *70*, pp. 663-674.

Li, Y. J., Marga, J. J., Cheung, C. M., Shen, X. L., & Lee, M. (2022). Health Misinformation on social media: A Systematic Literature Review and Future Research Directions. *AIS Transactions on Human-Computer Interaction*, vol. *14*(2), pp. 116-149.

Maathuis, C., Pieters, W., & van den Berg, J. (2018a). Developing a cyber operations computational ontology. Journal of Information Warfare, vol. 17(3), pp. 32-49.

Maathuis, C., Pieters, W., & van den Berg, J. (2018b). A knowledge-based model for assessing the effects of cyber warfare. In *Proceedings of the 12th NATO Conference on Operations Research and Analysis*.

Maathuis, C., & Chockalingam, S. (2022a). Responsible Digital Security Behaviour: Definition and Assessment Model. In *European Conference on Cyber Warfare and Security* (Vol. 21, No. 1).

Maathuis, C., & Chockalingam, S. (2022b). Victim versus Offender: Behaviour Modelling during Covid-19 Pandemic Cyber Attacks.

Maathuis, C. (2022a). An Outlook of Digital Twins in Offensive Military Cyber Operations. ECIAIR 2022. In *European Conference on the Impact of Artificial Intelligence and Robotics*, vol. 4, no. 1.

Maathuis, C. (2022b). On the Road to Designing Responsible AI Systems in Military Cyber Operations. In *European Conference on Cyber Warfare and Security* (Vol. 21, No. 1, pp. 170-177).

Mashiane, T., & Kritzinger, E. (2018). Cybersecurity behaviour: A conceptual taxonomy. In *IFIP International Conference on Information Security Theory and Practice* (pp. 147-156).

Mavrodieva, A. V., Rachman, O. K., Harahap, V. B., & Shaw, R. (2019). Role of social media as a soft power tool in raising public awareness and engagement in addressing climate change. *Climate*, vol. *7*(10), pp. 122.

Mazarr, M. J., Casey, A., Demus, A., Harold, S. W., Matthews, L. J., Beauchamp-Mustafaga, N., & Sladden, J. (2019). Hostile social manipulation present realities and emerging trends.

Mridha, M. F., Keya, A. J., Hamid, M. A., Monowar, M. M., & Rahman, M. S. (2021). A Comprehensive Review on Fake News Detection with Deep Learning. *IEEE Access*.

Marcot, B. G., & Penman, T. D. (2019). Advances in Bayesian network modelling: Integration of modelling technologies. *Environmental modelling & software*, vol. *111*, pp. 386-393.

Moran, G., Muzellec, L., & Johnson, D. (2019). Message content features and social media engagement: evidence from the media industry. *Journal of Product & Brand Management*.

NATO (2020). NATO's approach to countering disinformation: a focus on Covid-19. https://www.nato.int/cps/en/natohq/177273.htm

Nielsen, L. (2015). The Encyclopedia of human-computer interaction. 2nd ed. Aarhus, Denmark: The Interaction Design Foundationn; 2015 [Chapter 30]. Available from: https://www.interaction-design.org/encyclopedia/personas.html.

Pavlik, J. (2022). The Russian War in Ukraine and the Implications for the News Media. *Athens Journal of Mass Media and Communications*.

Pennycook, G., & Rand, D. G. (2021). The psychology of fake news. *Trends in cognitive sciences*, vol. *25*(5), pp. 388-402.

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of management information systems*, vol. *24*(3), pp. 45-77.

Peffers, K., Tuunanen, T., & Niehaves, B. (2018). Design science research genres: introduction to the special issue on exemplars and criteria for applicable design science research.

Perski, O., Blandford, A., West, R., & Michie, S. (2017). Conceptualising engagement with digital behaviour change interventions: a systematic review using principles from critical interpretive synthesis. *Translational behavioral medicine*, vol. *7*(2), pp. 254-267.

Petratos, P. N. (2021). Misinformation, disinformation, and fake news: Cyber risks to business. *Business Horizons*, vol. *64*(6), pp. 763-774.

Reisach, U. (2021). The responsibility of social media in times of societal and political manipulation. *European journal of operational research*, vol. *291*(3), pp. 906-917.

Schreiner, M., Fischer, T., & Riedl, R. (2021). Impact of content characteristics and emotion on behavioral engagement in social media: literature review and research agenda. *Electronic Commerce Research*, vol. *21*(2), pp. 329-345.

Traberg, C. S., Roozenbeek, J., & van der Linden, S. (2022). Psychological inoculation against misinformation: Current evidence and future directions. *The ANNALS of the American Academy of Political and Social Science*, vol. *700*(1), pp. 136-151.

Van Der Gaag, L. C. (1996). Bayesian belief networks: odds and ends. *The Computer Journal*, vol. *39*(2), pp. 97-113.

Wu, L., Morstatter, F., Carley, K. M., & Liu, H. (2019). Misinformation in social media: definition, manipulation, and detection. *ACM SIGKDD Explorations Newsletter*, vol. *21*(2), pp. 80-90.