8

SPECIAL TOPIC

Malevolent creativity and civil security:

The ambivalence of emergent technologies

Malevolente Kreativität und zivile Sicherheit:
Die Ambivalenz neu entstehender Technologien

Edited by A. Gazos, O. Madeira, G. Plattner, T. Röller, and C. Büscher

This article is licensed under a Creative Commons Attribution 4.0 International License (CC BY).
https://doi.org/10.14512/tatup.33.2.08

TATuP Zeitschrift für Technikfolgenabschätzung in Theorie und Praxis
Journal for Technology Assessment in Theory and Practice

oekom science

INTRODUCTION

New and emerging perspectives for technology assessment: Malevolent creativity and civil security

Alexandros Gazos[1], Octavia Madeira[1], Georg Plattner*[,1], Tim Röller[1], Christian Büscher[1]

Abstract • The characteristics of new technologies can cause significant harm to society. This does not only apply to unintended consequences, but even more so when the technologies are used for malicious purposes. The latter can be observed in the domain of *civil security*. Here, negative developments such as social polarization, increasing radicalization, and democratic regression alongside a deteriorating security situation are increasingly associated with technological innovation and *malevolent creativity*. As society becomes more and more aware of such correlations, technology assessment is called upon to provide orientation knowledge. This requires appropriate approaches and knowledge tools to assess the potential role of technology in relation to radicalization and political extremism.

***Neue und künftige Perspektiven für die Technikfolgenabschätzung**: Malevolente Kreativität und zivile Sicherheit*

Zusammenfassung • *Die Eigenschaften neuer Technologien können der Gesellschaft erheblichen Schaden zufügen. Dies gilt nicht nur für unbeabsichtigte Folgen, sondern vor allem dann, wenn die Technologien für böswillige Zwecke eingesetzt werden. Letzteres ist im Bereich der zivilen Sicherheit zu beobachten. Hier werden negative Entwicklungen wie soziale Polarisierung, zunehmende Radikalisierung und demokratischer Rückschritt sowie eine sich verschlechternde Sicherheitslage zunehmend mit technologischer Innovation und malevolenter Kreativität in Verbindung gebracht. Da sich die Gesellschaft solcher Zusammenhänge mehr und mehr bewusst wird, ist die Technikfolgenabschätzung gefordert, Orientierungswissen bereitzustellen. Dazu bedarf es geeigneter Ansätze und Wissensinstrumente, um die potenzielle Rolle der Technik in Bezug auf Radikalisierung und politischem Extremismus zu bewerten.*

Keywords • *malevolent creativity, extremism, radicalization, technology assessment, civil security*

This article is part of the Special topic "Malevolent creativity and civil security: The ambivalence of emergent technologies," *edited by A. Gazos, O. Madeira, G. Plattner, T. Röller, and C. Büscher. https://doi.org/10.14512/tatup.33.2.08*

Introduction

The current global decline of democracy (Papada et al. 2023) has raised significant challenges to society. These challenges are partly exacerbated by new technologies, particularly in the realms of communication via technical platforms and artificial intelligence (AI). Social media allows for the radicalization of vulnerable individuals without the need for physical support structures or networks (Calvert 2024; Ware 2023). An example of this is the social media campaign that led thousands of European Muslims to join ISIS (Gates and Podder 2015). The radicalization of some far-right 'single perpetrators' is another example (Mølmen and Ravndal 2023). New technologies such as artificial intelligence are currently developing rapidly and could become appealing to extremists. There are already signs of this in the use of deepfakes. The uncertainty they introduce can make people vulnerable to misinformation and radicalize existing views (Nieweglowska et al. 2023). Raising awareness of technology-related risks and unintended and undesirable consequences of technology has been a concern of technology

* *Corresponding author: georg.plattner@kit.edu*
[1] *Institute for Technology Assessment and Systems Analysis, Karlsruhe Institute of Technology, Karlsruhe, DE*

© 2024 by the authors; licensee oekom. This Open Access article is licensed under a Creative Commons Attribution 4.0 International License (CC BY).
https://doi.org/10.14512/tatup.33.2.09
Received: 29. 04. 2024; revised version accepted: 13. 05. 2024;
published online: 28. 06. 2024 (editorial peer review)

assessment (TA) since its inception. So far, however, the focus has tended to be on the opportunistic and (supposedly) benign intentions of scientists, engineers, entrepreneurs, or investors, with little awareness of the societal and environmental consequences (already in Baram 1973). The question of how technology might be deliberately used by malevolent actors to undermine civil security has rarely been addressed in TA research – sometimes as a by-product, sometimes in thematic niches. This usually involves dual-use technologies that can be used for both military and civilian uses (Mahfoud et al. 2018).

In this TATuP Special topic, we want to shed light on new and emerging technologies at the nexus of civil security and malevolent (mis)use. We reflect on the possibilities and limitations of assessing these technologies in terms of their potential for malevolent use, such as extremist radicalization and terrorist violence. The focus is on how malevolent use cases can be anticipated, prevented, or controlled.

Using technology in a function other than that intended by developers and with the intent to harm others is an inherently creative process.

The challenge for technology assessment

The research interest of TA is closely linked to society's existing perceptions and expectations of new technologies. These are often associated with an optimistic view of innovation, linking new technologies to social progress and hopes for an improvement in human living conditions (Bogner 2021). TA has specialized in discovering the blind spots in this optimistic view and corresponding innovation agendas (Baram 1973; Collingridge 1980; Sadowski 2015). While TA was initially limited to the anticipation of consequences by experts in the respective fields, the theories, methods, and practices have been greatly expanded in the direction of inclusive approaches in recent decades. Of special note are the integration of the stakeholder perspective on desired technology with undesired 'side' effects (Genus and Coles 2005; Schot and Rip 1997) and considerations of early co-design in the innovation process, which dominate current debates (Guston and Sarewitz 2002; Stilgoe et al. 2013).

These approaches do not apply well to the case of malevolent use of technology because TA does not have direct access to actors with such intentions (as in the case of military research). If we had this access, for example, via past offenses on record (Dessecker et al. 2024), we would most likely only get information about current but not future possibilities. This begs the question of how TA can properly observe such developments?

A first area of investigation is 'civil security'. While technical innovation is generally seen as a desirable deviation, civil security focuses on *undesirable deviations*. The aim is to prevent the spread of extremist ideas and ideologies as well as violent and terrorist acts. This perspective is based on the premise of preventive action and the logic of suspicion against individuals, groups, and organizations. It analyzes risk factors that increase the likelihood of malicious acts, such as psychological predisposition, social disadvantage, ideological imprint, etc. (Bröckling 2015). However, technology cannot simply be added as a further risk factor since technology is not only ubiquitously available but some technologies depend on, are connected to, or enable other technologies for a potentially unlimited number of applications. Above all, every technological innovation is a deviation from existing norms and knowledge. Under this premise, any technical progress would be generally suspected of being malevolently exploitable (Kusche and Büscher 2021).

Another pillar for future TA research is a branch of research dedicated to the phenomenon of malevolent innovation/actions. Using technology in a function other than that intended by developers and with the intent to harm others is an inherently creative process. Cropley, Kaufman and Cropley (2008) call this "malevolent creativity". They define it as a form of creativity that "is deemed necessary by some society, group, or individual to fulfill goals they regard as desirable, but has serious negative consequences for some other group, these negative consequences being fully intended by the first group" (Cropley et al. 2008, p. 106). When we talk about 'malevolent actors,' we mean those whose actions intentionally harm or negatively affect other groups or individuals (Schmid 2013). The fact that technologies can become the object of malevolent creativity is related to their common openness to unintended uses. Several bodies of literature consequently address the relationship between innovation processes and the intentional use of technology for malevolent ends. The focus of research has been on issues such as the intent of persons or groups (Gill et al. 2013; McLaren 1993), innovation dynamics between extremist or terrorist actors and organizations responsible for civil security (Dolnik 2007; Jackson 2001), innovation processes in society at large, and the general availability of technology (Cronin 2020).

All these perspectives provide information about the factors that play a role in the innovation process. However, they do not allow us to say anything about the future use of a particular technology. Every technology is the successful simplification in the medium of causality, i.e., the creation of effects through the rig-

 https://doi.org/10.14512/tatup.33.2.09

orous selection of useful causal relationships. If a technology is to be functional, all other cause-effect-relationships, which are still effective, must prove to be of no effect to this system (Luhmann 2005). This applies to mechanical installations as well as algorithms. Innovations also often arise in experiments or niches, i.e., in situations where disruptive influences can be temporarily ignored (de Haan and Rotmans 2011). Moreover, as we know very well today, technology always works most effectively when used without consideration (and often: without responsibility) for all other influences on the natural environment, health, ethics, and society. This is often the case when used for economic and military purposes and, even more so, when used with malevolent intentions, such as terrorist actors indiscriminately harming people or protected goods. We also know very well that the possibilities offered by technologies only reveal themselves at a much later stage of their development. The *Tor Project* is a prominent example of such changes in use. Originally conceived as a "free haven" (Moore and Rid 2016, p. 16) for military personnel to communicate anonymously on the internet, it has evolved into a decentralized, anonymized, and publicly accessible network. Over time, this led to the emergence of hidden websites and services and eventually to the emergence of a darknet that can be used for nefarious activities (Chertoff 2017).

The psychologist Gibson (2015) addressed this issue by developing the concept of affordances to refer to the possibilities that nature or technical artifacts present to an observer. It was important to him to emphasize the relationship between the object and the observer so that the potentials are not solely attributed to one or the other. Potentials remain unused if they are not discovered. Discovery does not depend on obvious properties. The tinkerers and creative minds keep trying until a useful causal isolation can be realized. What this is used for remains open: "All these benefits and injuries, these safeties and dangers, these positive and negative affordances are properties of things taken with reference to an observer but not properties of the experiences of the observer" (Gibson 2015, p. 137). The concept of technical affordances is of analytical value for the further development of TA concepts because it sensitizes our research to the indeterminacy of technical possibilities in relation to social developments such as political conflict. This raises the following questions for our special topic: What do material and immaterial artifacts allow actors to do? How can we explore unforeseen possibilities in the future use of known and yet unknown technologies by known and unknown actors – possibilities that exceed original intentions? In other words, how can we gain some traction in the face of this factually, temporally, and socially boundless complexity?

Outlook for technology assessment

Given the high degree of ignorance about technological affordances, there is a need for continuous monitoring (see also Stirling 2010). This is especially true in the context of radicalization, extremism, and terrorism. Such monitoring cannot be achieved by individual projects but must be driven by a large number of different observers. In an ongoing research project, we have approached the complexity of this issue step by step, moving from a coarse to a more finely tuned radar through various selections. First, we systematically interviewed experts on the future use of technology by extremist and terrorist actors as part of a two-stage Delphi survey. The results of this survey were used to prioritize detailed analyses, with functionality and availability serving as selective criteria. For example, one particularly relevant technology identified in this analysis is machine learning, which can be used in various areas such as improving translation software, disseminating propaganda, supporting transnational organizations, and creating deepfakes and social bots (Büscher et al. 2022). Second, an expert workshop on surveillance technologies was held to discuss the possibilities and limits of the use of technology and the conditions for intervention, particularly with regard to surveillance technologies for physical and digital locations. The focus was on questions of legal legitimacy, social desirability, and the possibilities and limitations of selective assessment with regard to the information burden and strategy development in policing extremism (Büscher et al. 2022). Third, in our subsequent analysis, we identified the metaverse and specific applications from the field of AI as particularly relevant. The affordances of these technologies can only be anticipated at this stage. In a vision assessment workshop held with experts in Karlsruhe, scenarios were developed on freedom and security in the metaverse in the context of extremist actors. Another expert workshop dealt with the use of AI applications in an extremist and preventive context (Madeira et al. 2023).

How can we explore unforeseen possibilities in the future use of known and yet unknown technologies by known and unknown actors?

These activities lead to interesting new avenues of research for TA. Looking closer at the preliminary broad sweep of topics, it is possible to derive individual studies on specific technologies. More importantly, the next step to take here is the devel-

https://doi.org/10.14512/tatup.33.2.09

opment and implementation of an approach to preventing malevolent use in the sense of responsible research and innovation and prevention by design. It is crucial for such an approach to take into account technological affordances while utilizing the potential of emerging technologies for the entire spectrum of prevention work, such as measures that focus on improving social conditions, reducing opportunities for misuse, engaging in early intervention, deradicalization, and rehabilitation (Schmid 2020). By networking with other interested groups such as non-governmental organizations (NGOs), mass media, and interest groups, long-term technology monitoring should be able to contribute to the prevention of malevolent use of new technologies.

Contributions in this Special topic

This Special topic addresses the challenges and issues outlined above in various ways. With regard to digital technologies and AI in particular, Tanja Sinozic-Martinez and Jutta Jahnel show how TA can help improve existing cultures of assessing digital technologies. By drawing on Mary Kaldor's work on security, they demonstrate the contribution TA can make to security studies by focusing specifically on human security and its very own 'security culture' in analyzing AI innovation. The article can be understood as a call to action for practitioners and academics in TA to pay more attention to issues of security in AI innovation processes in order to shape a culture that is mindful of the individual, human rights, and societal aspects of innovation. It adds a perspective to this Special topic that considers the predominantly state security or business-oriented approach to AI innovation (both of which also take into account certain types of malevolent creativity) but also shows that there is more at stake for a democratic society. The comparison to international regimes for the use of nuclear, biological, and chemical technology is apt, given that AI already has the potential to disrupt societies as a whole. But unlike the non-proliferation regimes for weapons of mass destruction, regimes for digital technologies must consider issues that Sinozic-Martinez and Jahnel refer to as "TA security culture".

In the subsequent article, Yannick Veilleux-Lepage and Don Rassler address the question of how and why terrorists use new technologies in innovative ways. They present a novel framework for analyzing emerging technologies that is based on four key factors to determine their potential for malevolent use. The research article analyzes how "disruptive," "dual-use," "democratized," and "diffused" characteristics of a technology can help in assessing its susceptibility for terrorist uses (4D framework). Their two examples of additive manufacturing and unmanned aerial systems show how malevolent creativity is central to the dynamics of terrorist innovation. However, the question of which technology might become the object of malevolent creativity is determined by external realities, such as (market) availability, pricing, or user friendliness. The framework introduced by Veilleux-Lepage and Rassler is an important new way of looking at emerging technologies through the lens of counterterrorism – a perspective that is still sorely lacking in technology assessment. The 4D framework offers great insights not only into how terrorists innovate but also into why terrorists (and by extension extremists) are more likely to use certain technologies maliciously than others. By applying the framework in the assessment of emerging technologies, TA can help prevent malevolently creative actors from using new technologies for their political goals.

In contrast, Dennis Klinkhammer takes a more hands-on-approach to demonstrate the potential of large language models (LLMs) to generate malicious content. Generative AI is becoming increasingly prevalent in society. However, soon after its emergence, more and more voices are pointing out the potential misuse of these tools and urging caution. The author shows how both direct and indirect manipulation can be used to get an LLM to generate malicious content aimed at the target group – possibly even against the security measures implemented. As a result, the production and distribution of misinformation, for example, could become much easier. It is therefore not surprising that this article is also a call for interdisciplinary cooperation to prevent malicious use cases and an important contribution to technology assessment.

It is crucial to take into account technological affordances while utilizing the potential of emerging technologies for the entire spectrum of prevention work.

As a challenge for international humanitarian law, Vasilios Tzoufis and Nikolaos Petropoulos turn their attention to the nexus of autonomous weapons systems and new forms of warfare. The issue of causation and guilt is central to their article: Who is ultimately responsible for any harm caused by an autonomous weapons system? Who will be criminally liable for any potential crime committed by these systems? These questions go to the core of international humanitarian law (IHL), and since autonomous weapons will change the very nature of war as we know it, the authors call on international actors to ensure that these new weapons will respect human rights in the future. Like other authors of the Special topic, Tzoufis and Petropou-

los call for international cooperation in this effort, with international uniform rules and protocols that set clear limits to the use of autonomous weapons systems in regular warfare. In this way, malicious and unlawful uses could be mitigated, at least in the arena of international relations, and, by extension, these international regimes could inform how innovators design their systems to comply with the norms of IHL.

International conventions to address the challenges posed by emerging technologies require national equivalents that take account of national specificities. Petr Machleidt, Jitka Mráčková, and Karel Mráček locate the need for AI regulation in relation to nationally nuanced risk perceptions by comparing (inter)national regulatory frameworks to Czech regulation. To find a holistic and measured approach to preventing and mitigating actual risks of malevolent AI use cases, the authors propose the application of the precautionary principle in regulation and education. The practice of TA could benefit from developing and implementing comprehensive approaches that address several challenges and research areas at the same time. Where these approaches lack clarity and focus, they can be situated in different socio-technical contexts. Situated approaches to holistic concepts can draw attention to commonalities, niches, and conflicting expectations. For example, the authors call for continuous assessment of responsibility and liability, as this would include the regulation of the entire value chain of emerging technologies, from development and innovation practices to use and misuse cases.

By the time technological innovations become relevant to the everyday life of users, their characteristics and thus also the fundamental opportunities for malicious use of technology have already been significantly shaped in previous innovation processes. Today, the innovation landscape and the ability to shape technologies is dominated by the private sector. The article by Niklas Henke presents a qualitative study that explores whether cross-sectoral practices for the development, variation, and anticipation of use cases can be found in private sector innovation projects and what significance these have in the development of new technologies in order to avoid undesirable use cases. Such explicit practices for the development of malevolent use cases in industrial innovation processes could be informative for the methodological development of a technology assessment interested in malevolent creativity, for example, through the identification of 'best practices.' Vice versa, the insights gained so far in TA on anticipating the malevolent potential of technology could also provide important impetus for innovation activities in the private sector. Opening up private sector innovation activities to the concerns of technology assessment is certainly desirable, if not necessary, against the background of the subject area dealt with in this Special topic.

The fact that this is not necessarily a futile endeavor and that market-oriented players may also have an interest in TA knowledge is demonstrated by undertakings such as the Immersive Democracy project. It is concerned with the question of how to safeguard democracy and democratic practices in an immersive future that is placed in the virtual world. Matthias Quent, head of the project, explains in his interview with the Special topic editors (see the Interview section) how they want to help shape the internet of the future and prevent society from making the same mistakes as in the Web 2.0 phase. The future of the Web might well be immersive, but equally important to the actual construction of a feasible metaverse will be questions about the safety of its users and its conduciveness for democracy.

Long-term technology monitoring projects can be an important tool to further develop analytical frameworks and apply them to new technologies at a stage where intervention is still possible.

Conclusion

This TATuP Special topic is conceived as an attempt to put issues of malevolent creativity and civil security on the map of TA. The result is a multidisciplinary look at various perspectives from which these issues can be discussed. The wide variety of contributions also shows that TA has much to offer for other fields of study, and other fields of study can offer much to TA. The topics range from philosophy to business and terrorism studies, from a theoretical focus to empirical data on how technology can be used by malevolent actors. Matthias Quent argues how democracy itself can be at stake if we do not look at these technological innovations through the lens of civil security.

Long-term technology monitoring projects can be an important tool to further develop analytical frameworks such as those presented in this Special topic and apply them to new technologies at a stage where intervention is still possible. The project Monitoring System and Transfer Platform Radicalization (MOTRA 2024) is the first to include such a dedicated module, at least in the German context, and the results show that such an approach offers valuable insights for state actors, prevention practitioners, and academia.

https://doi.org/10.14512/tatup.33.2.09

The apparent failure to anticipate the malevolent potential of Web 2.0, which has partly facilitated the decline of democracy seen around the world today, must not be repeated in the age of AI and immersive realities. And while malevolent creativity and use cases can never be completely prevented, it is still important to anticipate potential malevolent uses to keep technology, and by extension societies, as safe as possible.

Much more attention is needed to truly establish civil security and malevolent use of technologies as a central topic of TA, but we hope that with this Special topic the first step has been taken.

Funding • This research was supported by the German Federal Ministry of Education and Research within the framework of the program MOTRA (Grant No. MOTRA-13N15218).

Competing interests • The authors declare no competing interests.

References

Baram, Michael (1973): Technology assessment and social control. In: Science 180 (4085), pp. 465–473. https://doi.org/10.1126/science.180.4085.465

Bogner, Alexander (2021): Politisierung, Demokratisierung, Pragmatisierung. Paradigmen der Technikfolgenabschätzung im Wandel der Zeit. In: Stefan Böschen, Armin Grunwald, Bettina-Johanna Krings and Christine Rösch (eds.): Technikfolgenabschätzung. Handbuch für Wissenschaft und Praxis. Baden-Baden: Nomos, pp. 43–58. https://doi.org/10.5771/9783748901990-41

Bröckling, Ulrich (2015): Der präventive Imperativ und die Ökonomisierung des Sozialen. In: Public Health Forum 21 (4), pp. 29–31. https://doi.org/10.1016/j.phf.2013.09.003

Büscher, Christian et al. (2022): Trends der zukünftigen Technologienutzung im Kontext von Extremismus und Terrorismus. Erste Erkenntnisse aus dem MOTRA-Technologiemonitoring. In: Uwe Kemmesies et al. (eds.): MOTRA-Monitor 2021. Wiesbaden: BKA, pp. 248–281.

Calvert, Julius (2024): Engagement with radical propaganda drives cognitive radicalization. An analysis of a right-wing online ecosystem. In: Journal of Strategic Security 17 (1), pp. 24–30. https://doi.org/10.5038/1944-0472.17.1.2160

Chertoff, Michael (2017): A public policy perspective of the Dark Web. In: Journal of Cyber Policy 2 (1), pp. 26–38. https://doi.org/10.1080/23738871.2017.1298643

Collingridge, David (1980): The social control of technology. New York, NY: St. Martin's Press.

Cronin, Audrey (2020): Power to the people. How open technological innovation is arming tomorrow's terrorists. New York, NY: Oxford University Press.

Cropley, David; Kaufman, James; Cropley, Arthur (2008): Malevolent creativity. A functional model of creativity in terrorism and crime. In: Creativity Research Journal 20 (2), pp. 105–115. https://doi.org/10.1080/10400410802059424

de Haan, Hans; Rotmans, Jan (2011): Patterns in transitions. Understanding complex chains of change. In: Technological Forecasting and Social Change 78 (1), pp. 90–102. https://doi.org/10.1016/j.techfore.2010.10.008

Dessecker, Axel; Fecher, Lena; Hirth, Maria-Anna; Hofmann, Rebecca; Muzaqi, Lavdim (2024): Medien- und Technologienutzung durch jihadistische Straftäter*innen. Wiesbaden: MOTRA-Verbund. https://doi.org/10.57671/MOTRA-2024002

Dolnik, Adam (2007): Understanding terrorist innovation. Technology, tactics and global trends. London: Routledge.

Gates, Scott; Podder, Sukanya (2015): Social media, recruitment, allegiance and the Islamic State. In: Perspectives on Terrorism 9 (4), pp. 107–116. Available online at https://pt.icct.nl/article/social-media-recruitment-allegiance-and-islamic-state, last accessed on 02. 05. 2024.

Genus, Audley; Coles, Anne-Marie (2005): On constructive technology assessment and limitations on public participation in technology assessment. In: Technology Analysis & Strategic Management 17 (4), pp. 433–443. https://doi.org/10.1080/09537320500357251

Gibson, James (2015): The ecological approach to visual perception. New York, NY: Psychology Press. https://doi.org/10.4324/9781315740218

Gill, Paul; Horgan, John; Hunter, Samuel; Cushenbery, Lily (2013): Malevolent creativity in terrorist organizations. In: The Journal of Creative Behavior 47 (2), pp. 125–151. https://doi.org/10.1002/jocb.28

Guston, David; Sarewitz, Daniel (2002): Real-time technology assessment. In: Technology in Society 24 (1–2), pp. 93–109. https://doi.org/10.1016/S0160-791X(01)00047-1

Jackson, Brian (2001): Technology acquisition by terrorist groups. Threat assessment informed by lessons from private sector technology adoption. In: Studies in Conflict & Terrorism 24 (3), pp. 183–213. https://doi.org/10.1080/10576100151130270

Kusche, Isabel; Büscher, Christian (2021): Technologiemonitoring zur Prävention von Extremismus und terroristischer Gewalt. In: Birgit Blättel-Mink (ed.): Gesellschaft unter Spannung. Verhandlungen des 40. Kongresses der Deutschen Gesellschaft für Soziologie 2020. Available online at https://publikationen.soziologie.de/index.php/kongressband_2020/article/view/1307, last accessed on 02. 05. 2024.

Luhmann, Niklas (2005): Risk. A sociological theory. New Brunswick, NJ: Aldine Transactions.

Madeira, Octavia; Plattner, Georg; Gazos, Alexandros; Röller, Tim; Büscher, Christian (2023): Technologiemonitoring. Das Potenzial von Metaverse und KI für extremistische Verwendungszwecke. In: Uwe Kemmesies et al. (eds.): Motra-Monitor 2022. Wiesbaden: BKA, pp. 226–252. Available online at https://www.motra.info/motra-monitor-2022/, last accessed on 02. 05. 2024.

Mahfoud, Tara; Aicardi, Christine; Datta, Saheli; Rose, Nikolas (2018): The limits of dual use. In: Issues in Science and Technology 34 (4), pp. 73–78. Available online at https://issues.org/the-limits-of-dual-use/https://www.motra.info/motra-monitor-2022/, last accessed on 02. 05. 2024.

McLaren, Robert (1993): The dark side of creativity. In: Creativity Research Journal 6 (1–2), pp. 137–144. https://doi.org/10.1080/10400419309534472

Mølmen, Guri; Ravndal, Jacob (2023): Mechanisms of online radicalisation. How the internet affects the radicalisation of extreme-right lone actor terrorists. In: Behavioral Sciences of Terrorism and Political Aggression 15 (4), pp. 463–487. https://doi.org/10.1080/19434472.2021.1993302

Moore, Daniel; Rid, Thomas (2016): Cryptopolitik and the darknet. In: Survival 58 (1), pp. 7–38. https://doi.org/10.1080/00396338.2016.1142085

MOTRA (2024): Radicalization Monitoring System and Transfer Platform. Available online at https://motra.info/en/motra-im-profil/projektbeschreibung/, last accessed on 02. 05. 2024.

Nieweglowska, Maja; Stellato, Cal; Sloman, Steven (2023): Deepfakes. Vehicles for radicalization, not persuasion. In: Current Directions in Psychological Science 32 (3), pp. 236–241. https://doi.org/10.1177/09637214231161321

Papada, Evie et al. (2023): Defiance in the face of autocratization. Democracy report 2023. Gothenburg: University of Gothenburg. https://dx.doi.org/10.2139/ssrn.4560857

Sadowski, Jathan (2015): Office of technology assessment. History, implementation, and participatory critique. In: Technology in Society 42, pp. 9–20. https://doi.org/10.1016/j.techsoc.2015.01.002

Schmid, Alex (2013): The definition of terrorism. In: Alex Schmid (ed.): The Routledge handbook of terrorism research. London: Routledge, pp. 39–157.

Schmid, Alex (2020): Terrorism prevention. Conceptual issues (definitions, typologies and theories). In: Alex Schmid (ed.): Handbook of terrorism prevention and preparedness. The Hague: ICCT Press, pp. 13–48. Available online at https://www.icct.nl/sites/default/files/2023-01/Chapter-2-Handbook-.pdfhttps://www.motra.info/motra-monitor-2022/, last accessed on 02. 05. 2024.

Schot, Johan; Rip, Arie (1997): The past and future of constructive technology assessment. In: Technological Forecasting and Social Change 54 (2–3), pp. 251–268. https://doi.org/10.1016/S0040-1625(96)00180-1

Stilgoe, Jack; Owen, Richard; Macnaghten, Phil (2013): Developing a framework for responsible innovation. In: Research Policy 42 (9), pp. 1568–1580. https://doi.org/10.1016/j.respol.2013.05.008

Stirling, Andy (2010): Keep it complex. In: Nature 468 (7327), pp. 1029–1031. https://doi.org/10.1038/4681029a

Ware, Jacob (2023): The third generation of online radicalization. Washington, DC: George Washington University.

ALEXANDROS GAZOS
is a member of the scientific staff at ITAS and a PhD student at KIT. He is currently working in the "MOTRA – Technology Monitoring (TM)" project. His research focuses on the resilience and vulnerabilities of critical infrastructures as well as the sociology of technology.

[Source: © Petra Kirchmer]

DR. OCTAVIA MADEIRA
is a postdoctoral researcher at ITAS/KIT. She is working in the MOTRA-TM project. Octavia holds a PhD in Psychology and her research interests include virtual reality, extremism/terrorism, and forensic psychology.

DR. GEORG PLATTNER
is a postdoctoral researcher at ITAS/KIT and a member of the MOTRA-TM project. He holds a PhD in Political Science from the University of Vienna. His research interests include radicalization and violent extremism, civil security, and party studies.

TIM RÖLLER
is a project assistant at ITAS/KIT and is working in the MOTRA-TM project. He holds a bachelor degree in Social Sciences.

DR. CHRISTIAN BÜSCHER
is a senior researcher at ITAS/KIT. He is the project leader of the MOTRA-TM project. His research topics cover the theoretical embedment of technology assessment, social science-based energy research and ecological sociology.

Call for Abstracts

"Beyond short-termism:
Strategies and perspectives for the long-term governance of socio-technical change"

TATuP Special topic in 34/2 (2025). Please send your submission by 12 August 2024

Please read the complete CfA at https://www.tatup.de

Guest editors of this TATuP Special topic:

Dr. Stefania Sardo[1], Dr. Sophie Kuppler[1], and PD Dr. Dirk Scheer[1]

[1] Institute for Technology Assessment and Systems Analysis, Karlsruhe Institute of Technology

TATuP Zeitschrift für Technikfolgenabschätzung in Theorie und Praxis
Journal for Technology Assessment in Theory and Practice

oekom science

RESEARCH ARTICLE

TA for human security: Aligning security cultures with human security in AI innovation

Tanja Sinozic-Martinez*,[1], Jutta Jahnel[2]

Abstract • This research article addresses the growing concerns about digital technologies and artificial intelligence and their impact on human security. It explores the inadequacies of current digital technology regulation in protecting fundamental human rights. The authors present a typology of three "security cultures for digital technology" based on Mary Kaldor's work on human security, linking international relations and critical security studies with technology assessment (TA). The following cultures are distinguished: dual use, cybersecurity, and TA. The article concludes with a call for collaborative efforts among policy makers, industry, and civil society to prioritize a human-centered approach and global cooperation mechanisms and to strengthen the TA culture in order to address AI innovation without compromising human rights.

***TA für menschliche Sicherheit**: Ausrichtung der Sicherheitskulturen auf die menschliche Sicherheit bei KI-Innovationen*

Zusammenfassung • *Dieser Forschungsartikel befasst sich mit den zunehmenden Bedenken hinsichtlich digitaler Technologien und künstlicher Intelligenz sowie deren Auswirkungen auf die menschliche Sicherheit. Er untersucht die Unzulänglichkeiten der derzeitigen Regulierung digitaler Technologien beim Schutz grundlegender Menschenrechte. Die Autorinnen stellen eine Typologie von drei „Sicherheitskulturen für die digitale Technologie" vor, die auf Mary Kaldors Arbeit zu menschlicher Sicherheit basiert und die Disziplin internationale Beziehungen und Critical Security Studies mit Technikfolgenabschätzung (TA) verbindet. Folgende Kulturen werden unterschieden: Dual Use, Cybersecurity und TA. Der Beitrag schließt mit einem Aufruf zur Zusammenarbeit zwischen politischen Entscheidungsträger*innen, Branchenführer*innen und der Zivilgesellschaft, um einen menschenzentrierten Ansatz und globale Kooperationsmechanismen in den Vordergrund zu stellen und die TA-Kultur zu stärken, damit KI-Innovationen ohne Beeinträchtigung der Menschenrechte angegangen werden können.*

Keywords • *security cultures, artificial intelligence, digital humanism, fundamental human rights, digital ethics*

This article is part of the Special topic "Malevolent creativity and civil security: The ambivalence of emergent technologies," *edited by A. Gazos, O. Madeira, G. Plattner, T. Röller, and C. Büscher. https://doi.org/10.14512/tatup.33.2.08*

Introduction

As of the current moment, there exists a state of concern surrounding the latest digital technology field, namely artificial intelligence (AI). This has led to a convergence of world leaders and prominent figures in the tech industry coming together at the 'AI Safety Summit' in the United Kingdom to deliberate on the profound risks associated with AI. In addition, the European Union (EU) is in the process of finalizing its AI Act, civil society is mobilizing to include foundational models in AI regulation, and the United States government is issuing an order aimed at regulating large language models (LLMs). Although an escalation in human security risks posed by increasing loss of privacy, addiction economy, and algorithmic bias are widely recognized, scientific evidence shows that digital technology regulation is unsuccessful in the protection of fundamental human rights. Global AI threats are 'new' and 'different' from those posed by nuclear, biological and chemical technologies, and dominant

* *Corresponding author: tanja.sinozic-martinez@wien.gv.at*
[1] *Magistratsdirektion, Bereichsleitung für Wissenschaft, Stadt Wien, Wien, AT*
[2] *Institute for Technology Assessment and Systems Analysis, Karlsruhe Institute of Technology, Karlsruhe, DE*

© 2024 by the authors; licensee oekom. This Open Access article is licensed under a Creative Commons Attribution 4.0 International License (CC BY).
https://doi.org/10.14512/tatup.33.2.16
Received: 05. 01. 2024; revised version accepted: 08. 04. 2024;
published online: 28. 06. 2024 (peer review)

OPEN ACCESS

security regimes are 'old' ways in dealing with them. This paper tries to make sense of these 'new' differences by developing a typology of three 'digital technology security cultures' to better understand the challenges of addressing the threats posed by AI-supported digital technology innovations to humans. We apply the terms 'security cultures' directly from Mary Kaldor's work on human security (Kaldor 2007). In developing the three cultural types, we link the discipline of international relations (IR) and its sub-field of critical security studies with the field of technology assessment (TA) for an exploration of the security dimensions of regulating digital technology for the purposes of human security and the protection of humanitarian space.

Kaldor uses the terms security culture to denote "a pattern of behavior, which includes norms and standards which correspond to a particular interpretation of security that is associated with a form of political authority or a set of power relations. A security culture comprises different combinations of ideas, rules, people, tools, tactics and infrastructure, linked to different types of political authority that come together to address or engage in large-scale violence" (Kaldor 2018, p. 2). For the purpose of this paper, security is used to draw attention to the violation of human rights such as privacy, freedom from discrimination, safety, freedom from cruel treatment, equality, democracy, social security, work protection, and education, through the design, manufacture and use of technological innovations using the example of AI-supported digital technologies. The objective is to frame AI innovation as part of an ecosystem for maintaining progress towards preserving and enhancing human life and wellbeing, while drawing attention to the structures and processes in society which ensure such progress. The structures and processes are here associated with a 'TA security culture' which is also the research gap that this theoretical contribution intends to fill. Our method relied upon a literature review of AI threats, security cultures and TA, with the aim of continuing existing discussions on the role of culture in addressing the ambivalence of emerging technologies. Culture is the unit of analysis, rather than technologies or sectors or products. By viewing TA through an IR lens, the dimension of global governance for human rights protection in the face of technological change can be more deeply investigated.

The following section develops the concept of digital technology security cultures, drawing the main inspiration from Mary Kaldor's work on security and linking it to the most recent steps in digital technology evolution, namely AI capabilities. Thereafter, the dual use security culture, which is historically based on international cooperation in the regulation of nuclear, biological and chemical weapons is described. A discussion on the cybersecurity security culture follows, which we differentiate from the dual use paradigm in its greater reliance on private firms and private investors. Thereafter, the TA security culture is explained, which we associated with a social and humanistic focus on individual rights protection in the face of digital innovation, but as lacking the governance frameworks and private sector support of the other two cultures. The final section concludes.

Digital technology security cultures

AI threats to humanity are currently a 'hot topic' in the literature (Werthner et al. 2024). The concentration of power in the hands of a few large firms, rapidly increasing capabilities to spread misinformation at scale and the challenges these processes create for democracy are counterpoints to the benefits of AI for progress in areas such as science and medicine. Since at least the 1960s scientific evidence has shown that science and technology does not cause societal ills such as bias and inequality, but that technology evolves through socio-technical and techno-economic processes and cultures, embedding itself within pathways which already exist. To counteract societal deficiencies such as environmental degradation and inequality, and to redirect technology away from being grafted on these existing deficiencies of socio-economic systems, it is helpful to look at the ways in which the current technological paradigm interacts with them. The phrase 'digital security cultures' refers to the various methods and strategies employed to safeguard people from digital threats. This is essential because the ongoing development of digital technology poses risks not only in military contexts but also in non-military settings, and these risks impact human safety and well-being. Kaldor in her description of the evolution of the concept of 'security' (defined more broadly than technology threats) following Buzan and Hansen (Kaldor 2018, p. 13; Buzan and Hansen 2009) state that although the term emerged to define threats of war and defense, more recently include societal cohesion and non-military threats and vulnerabilities. The term security can be used to denote objectives (such as safety), or practices and artefacts in use (such as scanners at airports) (Kaldor 2018, pp. 13–14). The UN Development Programme (UNDP 1994) report describes seven types of risks to security: economic, food, health, environmental, personal, community and political. Security cultures, following Kaldor, are non-national, they include strategy, focus on functions, and are defined by specific ways of exercising power (Kaldor 2018). A central feature of this conceptualization of cultures is transition and change. Kaldor (2018, p. 24) draws attention to the usefulness of looking at mechanisms through which cultures are reproduced, such as the market, events, public debates, media representations, professional career and training structures, to find out how to reverse what is happening that is worsening human security.

‚Digital technology' is used to mean a dynamically evolving bundle of products, services, knowledge, skills and resources originating in the telecommunications industries in the 1960s and 70s and associated with capabilities of high data processing power (Freeman 2009). The information and communication technologies (ICT) paradigm is characterized by changes in production, design, consumption, speed and ease of communication and networking, and interactions between people. Recent emphasized changes have been the use of robotics, biotechnology, and the Internet of Things. Further focus is placed on emerging products and services enabled through artificial intelligence and the related changes which occur in governance, power struc-

https://doi.org/10.14512/tatup.33.2.16

tures and human identity, drawing attention to threats to humanity and civility (Nida-Rümelin and Weidenfeld 2022; Werthner et al. 2022).

The risk of technology negatively affecting humanity is of course not a new concern. However, while propaganda, misinformation, precise drone strikes, addictive product features, and biased algorithms are old issues, they present impacts which go beyond regular technological advancements of automation, even though they share some similarities. A central concern is the extent and degree to which they conflict with and gradually erode the social, civil and human rights institutions which have taken a very long time to build (such as for example individual property rights, the right to privacy, and social security). In the previous four technological revolutions (Perez 2002) not as many institutions protecting human rights and social welfare existed as today. The year 2023 marks 75 years of the Universal Declaration of Human Rights. Children were not protected during the industrial revolution from doing hard labor, but after it. The social welfare state in the UK is a post-WWII phenomenon. The digital shifts that violate human rights are highly incremental (Frischmann and Selinger 2018), much more subtle than in previous technological revolutions, and therefore more difficult to reverse.

Dual use security culture

The perhaps oldest technology security culture has its roots in the Cold War era (although the concept of dual use is traceable back to the 17th century (McLeish and Nightingale 2007), and initially focused on the international governance of nuclear, chemical and biological weapons (Molas-Gallart and Robinson 1997) and can be called the 'dual use technology security culture'. The term 'dual use' was originally used to refer to technologies that can be applied for both military and civilian purposes (Alic 1994), but have evolved to include the concepts peaceful versus non-peaceful, legitimate versus illegitimate, and benevolent versus malevolent (Rath et al. 2014) technology use. An important institutionalization of the dual use technology security culture was established for nuclear weapons in the Treaty on the Non-Proliferation of Nuclear Weapons (NPT) (effective in 1970) and for chemicals with the Convention on the Prohibition of the Development, Production, Stockpiling and Use of Chemical Weapons and on their Destruction (CWC) (effective in 1997), emerging out of the Cold War era. The aim of CWC was to regulate "the development, production, acquisition, stockpiling, retention, transfer, and use of chemical weapons" (Tuerlings and Robinson n.d., p. 4) at the international level. The CWC, unlike the NPT which did not depend on a civilian industry, had a unique impact on private industry, namely the chemicals industry (Tuerlings and Robinson n.d.). The main challenge, as for digital technologies which are also, like chemicals, primarily a civilian industry, is in defeating the development of illegal weapons and the malevolent use of civilian products while maintaining their scientific (McLeish and Nightingale 2007), technical and civilian usefulness.

The dual use technology security culture for nuclear and chemical technologies was created by various actors from the civil sector, public policy makers, private actors, governments, NGOs, and, for chemicals in particular, the private sector. It took a period of 50–70 years to create agreements and treaties. Non-state actors and their networks played a crucial role in shaping this culture. The absence of a central agency, or global technology governance at the international level, and private industry backing, makes processes very slow. Further, McGrew (1992, p. 318) states that "international regime and institutions of global management tend to lack the authoritative means to ensure compliance with their decisions".

As technological change has taken place, the dual use culture has extended beyond traditional military applications and the potential threat of over-regulation to science (McLeish and Nightingale 2007), to include ICTs and neurotechnologies (Mahfoud et al. 2018). Armed drones have become prevalent in ongoing wars, as has the use of mobile phones as triggering devices (Kaldor 2018, p. 28), the use of social media for mobilization and spreading fear (Kaldor 2018, p. 28) and continuous surveillance is enabled by the use of virtually any ICT product and system. Especially content-generation techniques evolved as a powerful weapon for disinformation and deep fakes, where one can no longer rely on what we are seeing or hearing. These hybrid threats can disrupt public order, attack communication infrastructures and weaken democratic structures (Chesney and Citron 2019). Compared to nuclear weapons, the production of biological weapons is easy and inexpensive, and information on how to produce them is readily available (Selgelid 2009). Similarly, digital technologies rapidly become easier to misuse. This is in big contrast to nuclear science, where discoveries on weapons and illegitimate use were usually classified (Selgelid 2009). Following the three general laws of Isaac Asimov invented in the field of robot ethics the priority of tackling AI technologies should be based on the precondition that there will be no harmful effects for human beings even in the case of a malevolent misuse (Clarke 1994). Effective dual use governance requires collaboration among states, private sector, international organizations and civil society.

Cybersecurity security culture

‚Cybersecurity' is variably used to raise awareness to threats to nation states and individuals which are dependent on, or are enabled by, innovation in ICTs. The here so-called 'cybersecurity security culture' has originated in and is shaped by the US through a collaboration between the US government and US-based ICTs firms in Silicon Valley (Cavelty 2007). Included in pushing it forward are public and private sectors. It relies on technical fixes to 'cyber-threats', which are reliant upon innovation in the ICTs sector and include products and systems to re-

 https://doi.org/10.14512/tatup.33.2.16

duce harm from for example worms, viruses, hacking, malware and ransomware directed at information system infrastructures at the national level, as well as personal computers.

It is analyzed, amongst other disciplines and fields, in the IR literature in the sub-field of security studies. The cyber-threat narrative, which is shaped by public communications of the threats to national security from and to digital technology-based products and systems, originated in the US in the Reagan administration, and later on the military through the concept of information warfare, and was first institutionalized with the Computer Security Act in 1987. The US military-funded innovation response to the national threat narrative has since 1991 been focused on the development of products and systems such as surveillance, reconnaissance systems and long-range precision weapons systems (Cavelty 2007). The high public investment in data collection and weapons technologies continues a long historical tradition in military and public spending on private sector innovation (Mazzucato 2013). Private industry specializing in these products has become highly profitable (Penney et al. 2018). Cavelty (2007, p. 4) says that "cyber-threats have clearly not materialized as a 'real' national security threat". She argues that instead of focusing on the terms cyber and security, it is necessary to look at the narratives which produce threat indicators, who produces the knowledge and how, and how this shapes measures and policies (Dwyer et al. 2022, p. 9).

In addition to public investments to the private sectors for technological innovation for military purposes to mitigate the threats described in the 'threat narrative' of the US, a big part of the cybersecurity culture is the protection of individual computers from crimes which depend upon and attack digital infrastructures such as hacking, malware, and denial of service and crimes which occur in other domains as well such as financial fraud, phishing, pharming and extortion). AI technologies could also enhance cybercrime threats by malevolent impersonations or identity theft through deep fakes (Ciancaglini 2020). The main challenge for human rights protection is difficult to achieve under the leverage of the cybersecurity security culture because both the infrastructures and the solutions for protecting the infrastructures, systems and its components are reliant on the private sector (Bay 2016; Penney et al. 2018). Lindstrom (2012) says that although the dominant cybersecurity narrative focuses on risks to national security on a macro scale level, they are not necessarily the biggest risks, those are accrued by individuals on a micro scale level. Human rights and social protection are being compromised by digital technology firms and products, such as for example in the case of labor protection. The global 'laissez-faire' approach to the regulation of digital innovation has been criticized for being insufficient in protecting human rights, marginalizing civil society in this narrative. Cavelty, referring to the words of Coles-Kemp in a collaborative piece published in 2002, states that what is needed in security culture is "a positive change that benefits the security of people" (Dwyer et al. 2022, p. 18). This is where technology security cultures can benefit from the integration of the human rights focus of TA and related fields such as AI ethics, digital humanism and responsible innovation (RI).

Technology assessment security culture

By focusing on individual security through the strengthening of civil society and global collaboration (Kaldor 2007), the third technology security culture can be called the 'TA security culture'. The dominant narrative is about the threats to fundamental human rights such as privacy, autonomy, health and well-being perpetuated by technological innovation in specific products and systems. The TA security culture is created by communities of TA scholars and practitioners, scientists from disciplines such as AI ethics, philosophy, sociology, Science and Technology Studies (STS) and science communication, the public sector and civil servants. Apart from scientific research, activities include public engagement of science and technology, science communication, and policy advice. An important step in formalizing this culture in policy circles has been the inclusion of academics specializing in the ethics of AI and TA in ethics councils such as the German Ethics Council. The emerging social movement of digital humanism (Nida-Rümelin and Weidenfeld 2022; Werthner et al. 2019) is inclusive of different actors such as urban planning and administration, cultural institutions (such as museums) public decision-makers, private industry, scientists from disciplines such as computer science, philosophy, ethics, TA, sociology, political science, and business. Early work on the concept of digital humanism is associated with the Vienna Manifesto on Digital Humanism (Werthner et al. 2019), a document outlining core principles for digital progress in line with the protection and strengthening of fundamental human rights. The objective of this culture (which includes different scientific communities such as AI ethics, RI, STS and sociology of technology, among others) is to maintain human rights in the digital realm, rather than only criminalizing their violation (borrowing Kaldor's words in her description of human security and humanitarian space). The TA culture can be traced back to the Office of Technology Assessment and the Ethical Legal and Social Aspects method of technology assessment, and more recently the digital rights, RI, AI ethics and STS communities, and providing reflections on the compromising of human rights, ethics and social institutions with the deployment and use of AI-enabled products.

The TA security culture depends upon government support and the public financing of research projects, social movements, education, designated organizations and research institutions, as well as the private sector in maintaining an ethical and human right centered approach to innovation. The threat narrative of this culture are violations of human rights at the micro, meso and macro levels. Risks at the micro level are associated with safety issues of deep learning technologies and data processing. This category includes prejudice against disabled people (Venkit et al. 2022), violation of privacy and intellectual property rights

infringement during training, and risks of using LLMs without understanding how they generate digital content (Taecharungroj 2023; Daws 2020). In addition, threats with regard to illegal criminal activities such as the generation of non-consensual pornography or abuse of malevolent impersonations play an important role. Besides single persons, also firms and organizations could be targeted (Ciancaglini 2020). The third risk category summarizes societal and systemic threats on the macro scale, for example political security issues posed by generative AI due to its high potential for disinformation and manipulating public opinion (Chesney and Citron 2019). AI technologies could be misused for propaganda, manipulation of elections or for stoking social unrest, political polarization or radicalization with the aim to undermine democracy. Finally, some experts are warning against an upcoming enhanced human intelligence or superintelligence that could not be controlled by humans and may pose an existential risk for humanity (Bostrom 2014).

The TA security culture does not yet have strong support from national governments or from private industry. Its main supporters are academia, research institutions directed at civil society, human rights organizations, and some parts of the public sector. As the EU AI Act comes into force in 2026, it can be expected that firms will engage more productively with standards such as Value-Based Engineering (Spiekermann and Winkler 2022) integrating ethical concerns with their innovation processes. However, the limited support from large firm oligopolies is a critical vulnerability. Nevertheless, there are echoes of this culture in employee movements in some of these large firms (for example, Google), and this can increase as tools are developed and firm ecosystems are shaped in the direction of the protection of human rights in the face of rapid AI advancements.

Conclusion: Navigating the nexus of security cultures, human rights, and AI

In analyzing the complex intersections between security cultures, human rights, and AI our paper has identified critical issues in framing digital technologies as inclusive of threats to fundamental human rights, similarly to nuclear, biological and chemical technologies, and therefore a relevant and 'new' security issue. As AI-enabled products and systems rapidly diffuse, it becomes evident that safeguarding human rights necessitates a human security focused approach to technological innovation.

Our examination reveals three technology security cultures: dual use, cybersecurity, and TA. Each culture reflects a unique set of norms, practices, and actors influencing the governance of AI-enabled digital technology. The dual use culture, rooted in Cold War dynamics, emphasizes international cooperation to regulate high-risk technologies. In contrast, the cybersecurity culture, pioneered by the US, relies on technical fixes and private sector engagement to counter digital threats. The TA culture, emerging from academic and ethical domains, places human rights at the forefront of technological development.

Each security culture is accompanied by its own threat narrative. The dual use culture contends with the complex task of governing dual-use technologies and preventing their malevolent use predominantly at the macro level. The cybersecurity culture grapples with the dual role of private sector entities in both protecting and potentially compromising individuals and firms. The TA culture focuses on safeguarding rights at different levels but faces challenges in garnering support from governments and large firms. To reverse what is happening in the compromising of human rights from rapid datafication and diffusion of ICTs and AI-enabled products and systems, policy makers, industry leaders and civil society must collaborate to fill the gap of a human-centered approach in the two dominant technology security cultures. Governments need to, address structural issues contributing to human rights violations, and firms need to embed ethical considerations into their AI innovation processes as far as possible. Global cooperation mechanisms, similar to those established for the dual use of nuclear, chemical and biological technologies, need to be negotiated and implemented for digital technologies as Bostrom says "to shift world politics into a more harmonious register" (Bostrom 2014, p. 259). Civil society has a pivotal role in shaping technology governance, and initiatives that amplify the voices of advocacy groups, ethical scholars, and human rights organizations, need to be strengthened to provide a counterbalance to the influence of powerful oligopolies. Inclusion in AI development, which incorporates diverse perspectives from marginalized groups needs to be encourage to minimize algorithmic biases and improve the adaptability of AI to diverse cultural contexts. The strategic analysis presented in this paper underscored the centrality of human security and the importance of strengthening the TA security culture. Focused on social, environmental and ethical objectives, it offers a pathway to navigate the challenges posed by AI, support the beneficial and useful aspects of AI innovation, without compromising fundamental human rights.

Funding • This work has received no external funding.
Competing interests • The authors declare no competing interests.

References

Alic, John (1994): The dual use of technology. Concepts and policies. In: Technology in Society 16 (2), pp. 155–172. https://doi.org/10.1016/0160-791X(94)90027-2

Bay, Morten (2016): What is cybersecurity? In search of an encompassing definition for the post-Snowden era. In: French Journal for Media Research 6, pp. 1–28. Available online at https://frenchjournalformediaresearch.com:443/lodel-1.0/main/index.php?id=988, last accessed on 07.06.2024.

Bostrom, Nick (2014): Superintelligence. Paths, dangers, strategies. Oxford: Oxford University Press.

Buzan, Barry; Hansen, Lene (2009): The evolution of international security studies. Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/CBO9780511817762

Cavelty, Myriam (2007): Cyber-security and threat politics. US efforts to secure the information age. Abingdon: Routledge.

Chesney, Robert; Citron, Danielle (2019): Deep fakes. A looming challenge for privacy, democracy, and national security. In: California Law Review 107, pp. 1753–1819. https://doi.org/10.2139/ssrn.3213954

Ciancaglini, Vincenzo et al. (2020): Malicious uses and abuses of artificial intelligence. Available online at https://www.europol.europa.eu/cms/sites/default/files/documents/malicious_uses_and_abuses_of_artificial_intelligence_europol.pdf, last accessed on 25. 04. 2024.

Clarke, Roger (1994): Asimov's laws of robotics. Implications for information technology. In: Computer 27 (1), pp. 57–66. https://doi.org/10.1109/2.248881

Daws, Ryan (2020): Medical chatbot using OpenAI's GPT-3 told a fake patient to kill themselves. In: AI News, 28. 10. 2020. Available online at https://www.artificialintelligence-news.com/2020/10/28/medical-chatbot-openai-gpt3-patient-kill-themselves/, last accessed on 25. 04. 2024.

Dwyer, Andrew; Stevens, Clare; Muller, Lilly; Cavelty, Myriam; Coles-Kemp, Lizzie; Thornton, Pip (2022): What can a critical cybersecurity do? In: International Political Sociology 16 (3), pp. 1–26. https://doi.org/10.1093/ips/olac013

Freeman, Christopher (2009): The ICT paradigm. In: Robin Mansell et al. (eds.): The Oxford Handbook of Information and Communication Technologies. Oxford: Oxford University Press, pp. 34–54. https://doi.org/10.1093/oxfordhb/9780199548798.003.0002

Frischmann, Brett; Selinger, Evan (2018): Re-engineering humanity. Cambridge, UK: Cambridge University Press.

Kaldor, Mary (2007): Human security. Cambridge, UK: Polity.

Kaldor, Mary (2018): Global security cultures. Cambridge, UK: Polity.

Lindstrom, Gustav (2012): Meeting the cyber security challenge. In: GCSP Geneva Papers Research Series 7. Available online at https://www.files.ethz.ch/isn/147788/7-2012.pdf, last accessed on 25. 04. 2024.

Mahfoud, Tara; Aicardi, Christine; Datta, Saheli; Rose, Nikolas (2018): The limits of dual use. In: Issues in Science and Technology 34 (4), pp. 73–78. Available online at https://issues.org/the-limits-of-dual-use/, last accessed on 25. 04. 2024.

Mazzucato, Mariana (2013): The entrepreneurial state. Debunking public vs. private sector myths. London: Anthem Press.

McGrew, Anthony (1992): Global politics in a transitional era. In: Anthony McGrew and Paul Lewis (eds.): Global politics. Globalization and the Nation-State. Cambridge, UK: Polity, pp. 312–330.

McLeish, Caitriona; Nightingale, Paul (2007): Biosecurity, bioterrorism and the governance of science. The increasing convergence of science and security policy. In: Research Policy 36 (10), pp. 1635–1654. https://doi.org/10.1016/j.respol.2007.10.003

Molas-Gallart, Jordi; Robinson, Julian (1997): Assessment of dual technologies in the context of European security and defence. Report for the Scientific and Technological Options Assessment (STOA). Luxembourg: European Parliament.

Nida-Rümelin, Julian; Weidenfeld, Nathalie (2022): Digital humanism. For a humane transformation of democracy, economy and culture in the digital age. Cham: Springer. https://doi.org/10.1007/978-3-031-12482-2

Penney, Jon; McKune, Sarah; Gill, Lex; Deibert, Ronald (2018): Advancing human rights-by-design in the dual-use technology industry. In: Columbia Journal of International Affairs 71 (2), pp. 103–110. Available online at https://ssrn.com/abstract=3218975, last accessed on 07. 05. 2024.

Perez, Carlota (2003): Technological revolutions and financial capital. The dynamics of bubbles and golden ages. Cheltenham: Edward Elgar.

Rath, Johannes; Ischi, Monique; Perkins, Dana (2014): Evolution of different dual-use concepts in international and national law and its implications on research ethics and governance. In: Science and Engineering Ethics 20 (3), pp. 769–790. https://doi.org/10.1007/s11948-014-9519-y

Selgelid, Michael (2009): Governance of dual-use research. An ethical dilemma. In: Bulletin of the World Health Organization 87 (9), pp. 720–723. https://doi.org/10.2471/BLT.08.051383

Spiekermann, Sarah; Winkler, Till (2022): Value-based engineering with IEEE 7000. IEEE Technology and Society Magazine 41 (3), pp. 71–80. https://doi.org/10.1109/MTS.2022.3197116

Taecharungroj, Viriya (2023): "What can ChatGPT do?" Analyzing early reactions to the innovative AI chatbot on Twitter. In: Big Data and Cognitive Computing 7 (1), p. 35. https://doi.org/10.3390/bdcc7010035

Tuerlings, Emmanuelle; Robinson, Julian (n.d.): The trilateral network associated with the Chemical Weapons Convention. Case study for the UN vision project on global public policy networks. Sussex: University of Sussex. Available online at https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=6142cc246bd45db237f409d982a965f958fbbd1c, last accessed: 25. 04. 2024.

UNDP – United Nations Development Programme (1994): Human development report 1994. New York, NY: Oxford University Press. Available online at https://hdr.undp.org/system/files/documents/hdr1994encompletenostats.pdf, last accessed: 25. 04. 2024.

Venkit, Pranav; Srinath, Mukund; Wilson, Shomir (2022): A study of implicit bias in pretrained language models against people with disabilities. In: Nicoletta Calzolari et al. (eds.): Proceedings of the 29th International Conference on Computational Linguistics. Gyeongju: COLING, pp. 1324–1332. Available online at https://aclanthology.org/2022.coling-1.113.pdf, last accessed: 25. 04. 2024.

Werthner, Hannes et al. (2019): Vienna manifesto on digital humanism. Available online at https://caiml.org/dighum/dighum-manifesto/Vienna_Manifesto_on_Digital_Humanism_EN.pdf, last accessed: 25. 04. 2024.

Werthner, Hannes et al. (eds.) (2024): Introduction to digital humanism. A Textbook. Cham: Springer. https://doi.org/10.1007/978-3-031-45304-5

DR. TANJA SINOZIC-MARTINEZ
is expert in digital humanism at the City of Vienna director's office since 2024. While doing the research for this article, she was academy scientist at ITA, Vienna, and guest researcher at ITAS, Karlsruhe. Her educational background is in economics (LSE), and science and technology policy studies (SPRU, Sussex). Her work focuses on humanism in the innovation and regulation of AI.

DR. JUTTA JAHNEL
studied food chemistry and gained experiences at the interface between analytics, technology and law. Since 2010 she works as a researcher at ITAS, Karlsruhe. Her recent work focuses on risk governance of artificial intelligence. She led a project to advise the EU Parliament on how to deal with deepfakes in the new AI regulation.

https://doi.org/10.14512/tatup.33.2.16

TATuP Zeitschrift für Technikfolgenabschätzung in Theorie und Praxis
Journal for Technology Assessment in Theory and Practice

oekom science

RESEARCH ARTICLE

The paradox of progress: How 'disruptive,' 'dual-use,' 'democratized,' and 'diffused' technologies shape terrorist innovation

Don Rassler*,[1], Yannick Veilleux-Lepage[2]

Abstract • Building upon the existing literature on terror innovation, this research article introduces a new analytical framework that highlights four key elements of emerging technologies which facilitate their adoption by terrorist groups. We posit that technologies that are inherently 'disruptive,' 'dual-use,' 'democratized,' and 'diffused' are particularly susceptible to misuse by terrorists, and that it is the interplay between these factors that helps to drive future directions of tech-enabled terror. We draw on terror and extremist interest in and use of two key technology areas – unmanned aerial systems and additive manufacturing, specifically 3-D-printed firearms – as examples to highlight how the factors in the framework interact and how they help to drive innovative terror behavior. We conclude with a series of concise recommendations to mitigate harmful terrorist use of these and other disruptive technologies and limit the impact of future tech-enabled terrorist innovations.

***Das Paradox des Fortschritts**: Wie ‚disruptive', ‚Dual-Use-', ‚demokratisierte' und ‚verbreitete' Technologien die Innovation im Bereich des Terrorismus prägen*

Zusammenfassung • *Aufbauend auf der bestehenden Literatur zu Terrorinnovationen wird in diesem Forschungsartikel ein neuer Analyserahmen vorgestellt. Dieser hebt vier Schlüsselelemente emergenter Technologien hervor, die deren Übernahme durch terroristische Gruppen erleichtern. Wir gehen davon aus, dass Technologien, die von Natur aus ‚disruptiv', ‚dual-use', ‚demokratisiert' und ‚verbreitet' sind, besonders anfällig für terroristischen Missbrauch sind, und dass das Zusammenspiel dieser Faktoren die künftige Richtung des technologiegestützten Terrors mitbestimmen wird. Anhand des Interesses von Terroristen und Extremisten an zwei Schlüsseltechnologien – unbemannte Flugsysteme und additive Fertigung, insbesondere 3-D-gedruckte Schusswaffen – zeigen wir auf, wie die verschiedenen Faktoren des Analyserahmens zusammenwirken und wie sie dazu beitragen, innovatives Terrorverhalten zu fördern. Wir schließen mit einer Reihe von Empfehlungen, um die terroristische Nutzung dieser und anderer disruptiver Technologien einzudämmen und die schädlichen Auswirkungen künftiger technologiegestützter terroristischer Innovationen zu begrenzen.*

Keywords • *terrorist innovation, emerging technologies, unmanned aerial systems, additive manufacturing*

This article is part of the Special topic "Malevolent creativity and civil security: The ambivalence of emergent technologies," *edited by A. Gazos, O. Madeira, G. Plattner, T. Röller, and C. Büscher. https://doi.org/10.14512/tatup.33.2.08*

Introduction

Terrorist groups operate in inherently hostile environments. In these sorts of environments, they continuously strive to gain a competitive advantage against their adversaries, whether that be a state and its security apparatus, or other terrorist groups striving to appeal to the same constituency (Veilleux-Lepage et al. 2022). To gain such an advantage, violent non-state actors routinely experiment with new or emerging technologies (Cronin

* *Corresponding author: don.rassler@westpoint.edu*
[1] *Combating Terrorism Center, U.S. Military Academy at West Point, West Point, US*
[2] *Royal Military College of Canada, Kingston, CA*

© 2024 by the authors; licensee oekom. This Open Access article is licensed under a Creative Commons Attribution 4.0 International License (CC BY).
https://doi.org/10.14512/tatup.33.2.22
Received: 15. 01. 2024; revised version accepted: 08. 04. 2024;
published online: 28. 06. 2024 (peer review)

OPEN ACCESS

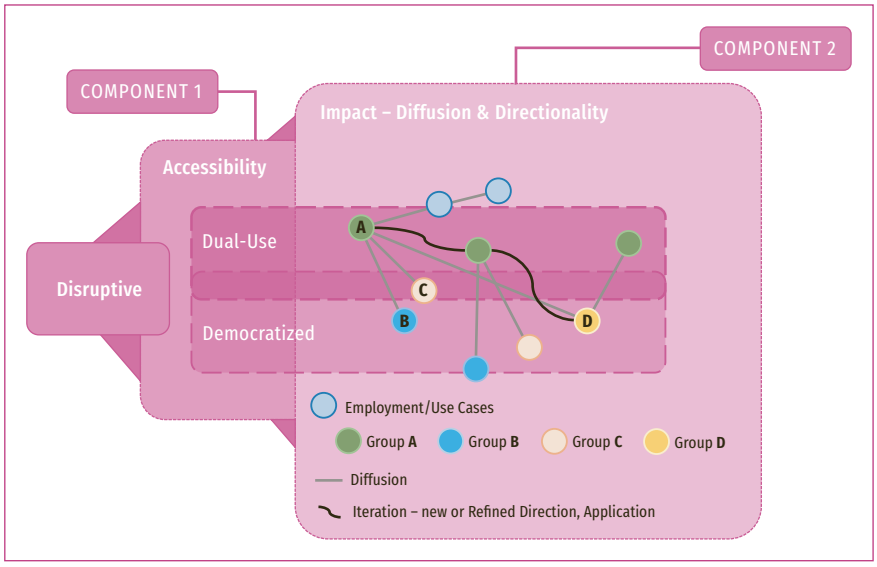


Figure 1: Framework – the four 'D' factors shaping terror innovation.
Source: authors' own compilation

2020). Two key disruptive technology areas that terror and extremist networks and movements have been heavily interested in, experimented with, and have used (or attempted to use) in innovative operational ways for the past decade, are unmanned aerial systems (UAS) and additive manufacturing specifically 3-D-printed firearms (3DPF) (Basra 2023; Chávez and Swed 2021; Rassler 2016; Veilleux-Lepage 2021; Veilleux-Lepage and Archambault 2022). As a result, these two technology areas provide important insights into how terrorists and extremists approach new technologies and the process and factors that help to shape innovative terror use of them. While terrorist groups must continuously innovate, the decision to leverage new technologies, nonetheless, presents an important paradox for terrorist organizations. Leveraging new technology is inherently a risky endeavor and doing so involves a cost-benefit judgment that might prove to be incorrect. Moreover, groups' efforts to adopt a new technology might fail for a variety of reasons and thus the group would pay the costs associated with experimenting with a new technology without reaping any of the anticipated benefits (Cragin et al. 2007; Veilleux-Lepage 2020). Terrorists and extremists also confront risks and costs when choosing not to adopt new technologies or innovate. They must find ways to strike a balance in making these decisions. This is because terror groups and movements who fail to innovate will either "be degraded to the point of irrelevance" or fail to attract resources, recruits and supporters (Gartenstein-Ross et al. 2019, p. 10). Building on the existing literature on terrorism innovation (Cragin et al. 2007; Dolnik 2007; Gartenstein-Ross and Joscelyn 2022; Lubrano 2021; Veilleux-Lepage 2020; Veilleux-Lepage et al. 2022), this article discusses how two key technologies – additive manufacturing and UAS – are currently or may in the future be leveraged by violent non-state actors by introducing a novel framework designed to showcase how the interplay and interrelationship between four 'D' factors – dual-use, democratized, diffuse, and directionality – help drive how extremists and terrorists use disruptive technologies. In doing so, this article highlights why these types of technologies, and other related technologies, warrant care and scrutiny. The article concludes with a series of short recommendations to mitigate harmful future terror use of key technologies and curb future, related terrorist innovations; these are aimed at law enforcement and intelligence agencies, domestic and international regulators, and commercial manufacturers.

Technology and the interplay of terror innovation

Terrorists and extremists can leverage a wide range of technologies to harm, achieve or enhance another desired effect, and/or to innovate. Disruptive technologies like additive manufacturing, UAS, and generative artificial intelligence, which have the potential to revolutionize entire industries or fields of science, and society generally, are particularly attractive as they hold the same potential and promise for extremists and terrorists too. As such, it is unlikely that these technologies will become contained. To use the words of noted Standford University bioethicists, "the centaur has left the barn" (Dowd 2005, para. 6).

The framework introduced in this article highlights how key accessibility factors interplay with one another and the influence and bearing these factors have on other important impact variables, which in turn help to drive both incremental and radical innovations by violent non-state actors. The framework is not designed to offer insights into *why* terrorists and extremists do or do not make innovative use of technologies, and the variables, tradeoffs, and decisions that violent non-state actors must navigate in evaluating whether to adopt a specific innovation or

technology. Instead, the framework offers a visual and conceptual perspective to help researchers and practitioners to understand how terror innovation and technology-related terror threats evolve as part of an interactive system. The framework, visually portrayed in Figure 1, is designed to showcase the interrelationship between four 'D' factors – dual-use, democratized, diffusion, and directionality – and how they interact and shape terror innovation.

The following two sections introduce the two main components of the framework. Examples of terror and extremist interest in and operational use of UAS and 3DPF are provided in both sections to highlight the applicability of the framework.

Accessibility – dual-use and democratized

The open-source technological revolution that Cronin (2020) describes is having a broadly disruptive impact on the tools, knowledge, and tradecraft available to state and non-state actors. The first component of the framework (Figure 1) highlights how this ongoing change in the accessibility of key technologies and information, and ways to share data, is helping state and non-state parties to enhance and diversify their capabilities, to develop new capabilities, and to innovate.

This change in accessibility has been principally shaped by the interplay between two key, interrelated factors: 1) dual-use items – products, software, and technologies designed for civilian purposes but can also be adapted for military use and 2) the democratized availability of systems and communication platforms that have made it easier and more efficient for individuals and networks to more securely share knowledge and technical know-how about how to use dual-use items, and to learn from other actors. Both factors continue to lower barriers to entry for non-state actors. Not only has this provided non-state actors with more tools and options to creatively tinker and experiment with, which opens pathways for them to innovate, it also has been shifting power from "dominant players" to other types of "players" (Cronin 2020, p. 7).

The Islamic State's innovative transformation of commercial-off-the-shelf (COTS) drones into airborne munitions-dropping weapons is an important example. For a considerable period, drones were primarily developed by states for military purposes. But, over the past two decades drone technology, and capable commercial UAS have become much more accessible and affordable for the average consumer. The Islamic State recognized this as an opportunity, and it was able to acquire a large fleet of small COTS drones through a global and layered supply chain (Rassler 2018). Then, through some creative tinkering the group quickly developed a fleet of small, easy-and-cheap-to-replicate munition-dropping capable drones; a Do-It-Yourself air force (Rassler 2018). For several months, the only effective solution to the Islamic State's new weaponized drones was small arms fire (Larter 2017).

While the Islamic State's creative weaponization of a dual-use item (i.e. commercial drones) led to its initial breakthrough innovation, the legacy of the group's innovation is tied to how its simple modifications – the know-how or playbook for how it weaponized its drones – was democratized, widely publicized, and shared around the world. Both elements, the use of a dual-use item and the democratized sharing of innovation know-how, and the interplay between those two factors, have facilitated the diffusion of munition-dropping commercial drones as a new method to attack for different types of actors. For example, a wide range of actors – from Mexican drug cartels to Iranian proxies and state militaries, have been inspired by, have leveraged, and in some cases have sought to improve upon and take in new directions, the Islamic State's 'drop-a-munition-from-a-commercial-drone' idea. The fact that the Ukrainian and Russian militaries have both used and experimented with munition-dropping drones similar to those used by the Islamic State (Axe 2022; Chapple 2022) illustrates that the learning ecosystem for how dual-use items can be used for harm cuts across states and non-state actors as distinct categories. It is clear that both 'types' of actors are learning lessons from each other, which helps to further diffuse innovative ideas and capabilities, and push those capabilities in new directions.

Over the past two decades drone technology, and capable commercial UAS have become much more accessible and affordable for the average consumer.

Terrorist interest in and attempts to use 3DPF (Veilleux-Lepage 2021) also highlights how dual-use items and democratized knowledge and know-how interact and can be combined to generate new pathways for mal-intended actors to inflict harm. For example, to create a 3DPF, an actor must first acquire or gain access to hardware – an actual 3D printer (i.e. a dual-use commercial good). Today, there are numerous high-quality 3D printers on the market for under $ 1,000 US. A second requirement is software that enables a 3D printer to 'read' a computer-animated design (CAD) file so it can 'print' or physically create the object. Software is thus a category that cuts across the dual-use and democratized categories as concepts, as it can be both a dual-use item (i.e. software that comes pre-loaded on a 3D printer) and can be acquired or shared in more decentralized ways. There is an entire online ecosystem where 3D printing hobbyists share design files and technical know-how, where

 https://doi.org/10.14512/tatup.33.2.22

technical information is democratized and shared. There are segments of that rich open-source community that support the development of 3DPF, and ensuring that individuals maintain access to information, plans, and software that would allow them independently to produce such weapons. Mainstream tech companies have banned the sharing of CAD files for 3DPF on their platforms (Barton 2019), but a litany of websites and chat servers emerged to fill the void left by these crackdowns (Sorensen 2019). At the same time, plans for 3DPF have evolved and become significantly more sophisticated since they were introduced online in 2013.

Important points of intersection between extremist and online 3DPF communities have emerged, and terrorist interest in the operational use of 3DPF has also increased across time. For example, according to research the creator of "the world's most popular 3-D-printed gun […] the FGC-9" posted extremist messages and "made anonymous threats of violence" online (Basra 2023, executive summary). From 2016 to December 2023, a total of 27 terrorist plots involving 3DPF were disrupted across various countries, illustrating the growing concern over this technology's use in terrorism, and its appeal. The United Kingdom reported the highest number of incidents, with seven cases, while the United States followed with three cases. Other countries also faced such threats.

The timeline of these incidents shows an increasing trend. In 2016, there was just one incident, and 2019 also saw a single case. However, the numbers rose in subsequent years, with three incidents in 2020, six in each 2021 and 2022, and a significant jump to ten incidents in 2023. Among these, some plots were particularly alarming. In October 2019, for example, Stephan Balliet conducted a terrorist attack in Halle, Germany, which he livestreamed. Balliet's arsenal of weapons included a longsword (not used during the attack), pipe bombs and improvised explosive devices, and six firearms – five of which were improvised. While these weapons were mostly constructed out of steel, aluminum, and wood, Balliet used some 3DPF components. According to his manifesto, one of his motivations for the attack was to "prove the viability of improvised weapons" (Koehler 2019, p. 15).

More recently, in 2022, Icelandic authorities arrested four individuals linked to a Neo Nazi terror plot. This group planned to use a combination of semi-automatic and 3DPF to launch attacks on various societal institutions (Boffey 2022). These arrests were part of a larger European operation, leading to police raids in Germany, Italy, Croatia, and Lithuania.

Impact – diffusion and directionality

The second component of the framework (also visually portrayed in Figure 1) involves the inclusion of two other 'Ds' – diffusion and directionality. Diffusion involves "dispersion through a space or over a surface." (OED 2023), whereas directionality can be understood as "relating to direction in space" (MWD 2024). While the previous two 'Ds' – dual-use and democratized – primarily affect the accessibility of equipment, technology, systems, and know-how that is available to state and non-state actors, the diffusion of key innovations, capabilities, and know-how helps to amplify their impact. This in turn – through further iteration by various parties – helps to push technology-related threats in different directions.

The nature of the threats posed by terrorists leveraging these emerging technologies cannot be fully understood by looking at developments in the terrorism sphere alone.

The framework attempts to visually capture the interplay between these four 'D' factors and how they interact and influence each other as part of a system, in two ways. First, is the build character of the framework itself. The two accessibility factors – dual-use and democratized – are core, foundational 'drivers' that guide and make the impact factors of diffusion and directionality more likely. Or put another way, the accessibility factors – due to their open nature – help ideas, tactics, and the development of new weapons or operational approaches to diffuse. This creates an environment where a wide range of actors (i.e., individuals, non-state actors, and states) have an expanded and more diversified pool of diffused materiel and know-how to pull from and to create; a dynamic which broadens opportunity pathways, or directions, an actor can push a threat. The simplest way to characterize the link between the accessibility and impact factors is that more diversified inputs will likely lead to more diversified outcomes.

The state and projections of the global drone market help to highlight these interplay dynamics. In 2021 the global market for COTS UAVs was valued at over 20.8 billion USD, and it is projected to further grow to nearly 1,205 billion USD by 2030 (Straits Research 2021). The global drone market is also "diverse and multifaceted, encompassing various segments" (Beekman 2023). Ongoing advancements and innovations occurring across a range of key technologies, from sensors to powering systems, and data processing capabilities, will also ensure that tomorrow's COTS UAVs will be more capable than they are today.

This means that not only will violent non-state actor use of drones remain a persistent problem and increase in scale, the character of the terror drone threat, and the ways in which it presents itself and directions it takes, will likely become more diversified as well. For example, the majority of COTS UAVs

https://doi.org/10.14512/tatup.33.2.22

available today are powered by lithium-ion batteries, which limit how far and for how long those drones can fly. When viewed through the lens of extremism and terrorism, this limits the range from which a terror entity can conduct a stand-off attack. But, "alternative powering options, such as UAS powered by hydrogen fuel cell technology or hybrid fuel/powering systems (i.e., solar), are already commercially available" (Rassler 2024, p. 5). In the next decade "hydrogen fuel cell and solar UAS technology will evolve and mature, and will also likely become more available to the average consumer, which will make longer ranges more accessible" (Rassler 2024, p. 5). This will open-up new range options, and help to push the terror drone threat in new directions.

The second important takeaway from the framework is that the interplay between the four 'Ds' is hard to box or bound. For example, dual-use and democratized are factors that have an ongoing influence and bearing on the diffusion, or spread, of new innovations and capabilities. As a result, dual-use and democratized are shown as factors that cut across the meta-accessibility and impact categories. The circular employment use cases, linear diffusion paths (blue lines), and curvy iteration courses (black lines) in Figure 1 have been included to help model in a general way how the 4 'D' factors interact and influence one another. For example, theoretical terror Group A (green circle) recognizes the broad availability of commercial UAV quadcopters and decides to acquire several stock quadcopters to conduct surveillance missions and/or to support propaganda. Groups B and C observe that Group A is utilizing commercial quadcopters in these ways, and decide to copy and mirror the idea, highlighting how knowledge related to the use of the dual-use good (i.e., COTS UAV) for those purposes democratized and the tactic diffused.

Group A then decides that it wants to weaponize a commercial quadcopter. As the first curvy black line highlights, Group A experiments until it develops a creative and low-cost way to drop munitions from a UAV; an innovation that pushes the threat in new directions. Groups B and C mirror this new 'drop a munition from a commercial UAV' approach too. Group D, which has been observing all of this activity, iterates on Group's A UAS weaponization approach and develops a way to integrate commercially available crowd counting software, which "utilizes deep learning technology to quickly and accurately analyze live video," (Canon USA 2024) so the group can maximize of the impact, and lethality, of its UAS bombs.

The findings from a comparative study of the drone programs of five-armed non-state actors also help to bring some of these diffusion and directionality considerations to light (Veilleux-Lepage and Archambault 2022). The study found that all five non-state entities used or developed weaponized drones, and that all five have used pilot-to-target as a drone attack modality. So, when looking across the five programs, pilot-to-target was a common drone weaponization feature: a diffused attack method. But that study also found that there were important differences in how each of the five non-state entities used drones. For example, while each of the five groups have used pilot-to-target as a drone weaponization pathway, how each of the five groups have utilized drones in that way has been guided by a group's own preferences (Veilleux-Lepage and Archambault 2023).

The fact that the study found "that there is no single route of development for the use of drones by non-state entities, nor is there a pattern that these groups want to follow in order to expand their capabilities" (Veilleux-Lepage and Archambault 2022, p. 1) highlights why directionality is also an important factor for the framework to consider. Indeed, since "each organization uses drones in a manner that is unique to its own set of logistical, political, and strategic parameters" (Veilleux-Lepage and Archambault 2022, p. 1) considering directionality as a key factor will help practitioners to better anticipate, and be on the look-out for, differences and change.

Even though how different non-state actors use drones, or other disruptive technologies, will vary from group to group or network to network, directionality will still be guided by several primary characteristics that aim to improve operational outcomes. These include characteristics such as surprise and speed, effectiveness and/or efficiency, as well as impact or outcome variables like lethality.

Responses and policy recommendations

Advancements in key technologies and their widespread availability offer significant benefits, yet they also pose risks by providing malevolent actors with new capabilities. The interaction of the four 'D' factors in our framework is a driving force behind the evolution of violent non-state threats.

Law enforcement and intelligence

If not already being done on a broad scale, governments should start systematically cataloging cases of proven interest in, and use of, UAS and additive manufacturing by terrorist actors, as well as innovative ways that private citizens have used these technologies. This data would allow those tracking these threats to quickly identify noteworthy developments and to better understand the evolution of these phenomena. Such cataloguing could initially be implemented and maintained with minimal resources and effort, and it would pay dividends in the years ahead as it would help government officials make more informed resourcing and regulatory decisions.

In addition, the nature of the threats posed by terrorists leveraging these emerging technologies cannot be fully understood by looking at developments in the terrorism sphere alone. For example, evaluating how terrorists might leverage additive manufacturing requires law enforcement and intelligence agencies to pay attention and understand innovative developments that occur in private industry, academia and by creatively minded hobbyists. To stay current, analysts following these issues therefore need to monitor innovations across a number of spheres, and be part terror specialist, part technologist, and part industry expert.

 https://doi.org/10.14512/tatup.33.2.22

The overwhelming majority of game-changing developments in these spheres is taking place in plain sight.

Multinational

A potentially fruitful avenue for international cooperation is the United Nations Security Council Resolution 2325. Unanimously approved, the resolution called upon states to take into account the risks of terrorist groups using "rapid advances in science, technology, and international commerce for proliferation purposes" (Black-Branch 2017, p. 228). In addition, the Resolution also added controlling access to intangible transfers of technology as a new obligation for UN member-states. The Resolution also recognizes the need to continue drawing upon relevant expertise from industry, the scientific community, and academia in order to ensure access to up-to-date information on these technologies and in order to educate these communities about the terrorism challenges at hand.

Commercial

Finally, manufacturers can help counter misuse of emerging technologies by adopting 'Design Against Crime' (DAC) practices. The DAC involves designing products in such a way as to reduce the harmful 'misuse' of emerging technologies (Ekbom 2005). For example, most modern commercially sold photocopiers will automatically recognize banknotes, and either decline to reproduce them or produce a distorted image of them to prevent counterfeiting (Girgensohn et al. 2018).

While promising, DAC initiatives require manufacturers to introduce such features in close coordination with both public authorities and also other manufacturers in order to ensure maximum reliance on common standards and protocols. In addition, manufacturers can also further protect their products from being misused by sharing technical information about future products with regulatory and government agencies.

Conclusion

This article delves into how additive manufacturing and UAS are being, or could be in the near future, exploited by violent non-state actors. It introduces a novel framework to analyze the dynamics of terrorists' use of disruptive technologies, focusing on four key factors: dual-use, democratized, diffused, and directionality. These factors not only drive extremists' use of these technologies but also underline why such technologies demand careful observation and regulation.

The framework presented elucidates how access to technology, shaped by its dual-use nature and democratized availability, enhances the capabilities of state and non-state actors alike. It underscores the lowering of barriers to entry for non-state actors, shifting power dynamics, and factors fostering innovation. Furthermore, the interplay of dual-use and democratized factors not only facilitates the diffusion of these technologies but also influences their directionality – the ways in which they are adapted and evolved for specific purposes. This aspect is vital for understanding and anticipating the evolution of technology-related terror threats.

In the face of these emerging threats, the article proposes targeted recommendations for law enforcement, intelligence agencies, international regulators, and commercial manufacturers. Systematic cataloging of terrorist use of these technologies, cross-sectoral monitoring, and innovative collaborations are crucial.

Funding • This work received no external funding.
Competing interests • The authors declare no competing interests.

References

Axe, David (2022): Ukraine's $ 10,000 drones are dropping tiny bombs on Russian troops. In: Forbes, 13. 04. 2022. Available online at https://www.forbes.com/sites/davidaxe/2022/04/13/ukraines-10000-drones-are-dropping-tiny-cheap-bombs-on-russian-troops/, last accessed on 11. 04. 2024.

Barton, Champe (2019): As social networks crack down, 3d-printed gun community moves to new platforms. In: The Trace 25. 07. 2019. Available online at https://www.thetrace.org/2019/07/3d-printed-guns-social-media-ban/, last accessed on 11. 04. 2024.

Basra, Rajan (2023): Behind the mask. Uncovering the extremist messages of a 3d-printed gun designer. London: International Centre for the Study of Radicalisation. Available online at https://icsr.info/wp-content/uploads/2023/10/ICSR-Report-Behind-the-Mask-Uncovering-the-Extremist-Messages-of-a-3D%E2%80%91Printed-Gun-Designer.pdf, last accessed on 11. 04. 2024.

Beekman, Johannes (2023): Skyward bound. Navigating the evolving terrain of the global drone market. In: IoT Marketing, 15. 12. 2023. Available online at https://iotmktg.com/skyward-bound-navigating-evolving-terrain-of-the-global-drone-market/, last accessed on 11. 04. 2024.

Black-Branch, Jonathan (2017): Nuclear terrorism by states and non-state actors. Global responses to threats to military and human security in international law. In: Journal of Conflict and Security Law 22 (2), pp. 201–248. https://doi.org/10.1093/jcsl/krx004

Boffey, Daniel (2022): Icelandic police arrest four people over alleged terror attack plans. In: The Guardian, 22. 9. 2022. Available online at https://www.theguardian.com/world/2022/sep/22/icelandic-police-arrest-four-people-over-alleged-terror-attack-plans, last accessed on 11. 04. 2024.

Canon USA (2024): Crowd people counter. Available online at https://www.usa.canon.com/shop/p/crowd-people-counter-version-1-0?color=Black&type=New, last accessed on 11. 04. 2024.

Chapple, Amos (2022): The drones of the Ukraine war. In: Radio Free Europe/Radio Liberty, 17. 11. 2022. Available online at https://www.rferl.org/a/ukraine-russia-invasion-drones-war-types-list/32132833.html, last accessed on 11. 04. 2024.

Chávez, Kerry; Swed, Ori (2021): The proliferation of drones to violent nonstate actors. In: Defence Studies 21 (1), pp. 1–24. https://doi.org/10.1080/14702436.2020.1848426

Cragin, Kim; Chalk, Peter; Daly, Sara; Jackson, Brian (2007): Sharing the dragon's teeth. Terrorist groups and the exchange of new technologies. Santa Monica, CA: RAND Corporation. https://doi.org/10.7249/MG485

Cronin, Audrey (2020): Power to the people. How open technological innovation is arming tomorrow's terrorists. Oxford: Oxford University Press.

https://doi.org/10.14512/tatup.33.2.22

Dolnik, Adam (2007): Understanding terrorist innovation. Technology, tactics and global trends. London: Routledge.

Dowd, Maureen (2005): What rough beast? In: The New York Times, 07. 05. 2005. Available online at https://www.nytimes.com/2005/05/07/opinion/what-rough-beasts.html, last accessed on 11. 04. 2024.

Ekblom, Paul (2005): Designing products against crime. In: Nick Tilley (ed.): Handbook of crime prevention and community safety. London: Routledge, pp. 203–244. https://doi.org/10.4324/9781315724393

Gartenstein-Ross, Daveed; Joscelyn, Thomas (2022): Enemies near and far. How jihadist groups strategize, plot, and learn. New York, NY: Columbia University Press. https://doi.org/10.7312/gart19524

Gartenstein-Ross, Daveed; Shear, Matt; Jones, David (2019): Virtual plotters, drones, weaponized AI? Violent non-state actors as deadly early adopters. s.l.: Valens Global. Available online at https://valensglobal.com/virtual-plotters-drones-weaponized-ai-violent-non-state-actors-as-deadly-early-adopters/, last accessed on 11. 04. 2024.

Girgensohn, Andreas; Wilcox, Lynn; Liu, Qiong (2018): Automatic rights management for photocopiers. In: Proceedings of the ACM Symposium on Document Engineering 2018. New York, NY: ACM, pp. 1–10. https://doi.org/10.1145/3209280.3209531

Koehler, Daniel (2019): The Halle, Germany, synagogue attack and the evolution of the far-right terror threat. In: CTC Sentinel 12 (11), pp. 14–20. Available online at https://ctc.westpoint.edu/halle-germany-synagogue-attack-evolution-far-right-terror-threat/, last accessed on 11. 04. 2024.

Larter, David (2017): SOCOM commander. Armed ISIS drones were 2016's 'most daunting problem'. In: Defense News, 16. 05. 2017. Available online at https://www.defensenews.com/digital-show-dailies/sofic/2017/05/16/socom-commander-armed-isis-drones-were-2016s-most-daunting-problem/, last accessed on 11. 04. 2024.

Lubrano, Mauro (2021): Navigating terrorist innovation. A proposal for a conceptual framework on how terrorists innovate. In: Terrorism and Political Violence Routledge 35 (2), pp. 248–263. https://doi.org/10.1080/09546553.2021.1903440

MWD – Merriam-Webster Dictionary (2024): "Directionality". Available online at https://www.merriam-webster.com/dictionary/directionality, last accessed on 15. 04. 2024.

OED – Oxford English Dictionary (2023): "Diffusion". Available online at https://www.oed.com/dictionary/diffusion_n?tab=factsheet#6760287, last accessed on 15. 04. 2024.

Rassler, Don (2016): Remotely piloted innovation. Terrorism, drones and supportive technology. West Point, NY: Combating Terrorism Center at West Point.

Rassler, Don (2018): The Islamic State and drones. Supply, scale, and future threats. West Point, NY: Combating Terrorism Center at West Point.

Rassler, Don (2024): Going the distance. The emergence of long-range stand-off terrorism. In: CTC Sentinel 17 (2), pp. 1–10. Available online at https://ctc.westpoint.edu/going-the-distance-the-emergence-of-long-range-stand-off-terrorism/, last accessed on 15. 04. 2024.

Sorensen, Zachary (2019): 3D printed guns and gun control. In: Washington University Political Review, 19. 04. 2019. Available online at https://www.wupr.org/2019/04/19/3d-printed-guns-and-gun-control/, last accessed on 15. 04. 2024.

Straits Research (2021): Commercial drone market growth, analysis. Report to 2022–2030. Pune: Straits Research. Available online at https://straitsresearch.com/report/commercial-drone-market, last accessed on 15. 04. 2024.

UNSCR – United Nations Security Council (2016): Resolution 2325. Non-proliferation of weapons of mass destruction. New York, NY: United Nations. Available online at http://unscr.com/en/resolutions/2325, last accessed on 15. 04. 2024.

Veilleux-Lepage, Yannick (2020): How terror evolves. The emergence and spread of terrorist techniques. Lanham, MD: Rowman & Littlefield.

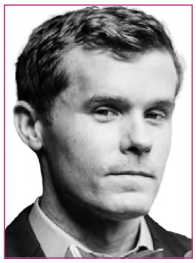
Veilleux-Lepage, Yannick (2021): Ctrl, hate, print. Terrorists and the appeal of 3-D-printed weapons. The Hague: International Centre for Counter-Terrorism – ICCT.

Veilleux-Lepage, Yannick; Archambault, Emil (2022): A comparative study of non-state violent drone use in the Middle East. The Hague: International Centre for Counter-Terrorism – ICCT.

Veilleux-Lepage, Yannick; Archambault, Emil (2023): Étude comparative de l'usage des drones par des groupes armés non étatiques au Moyen-Orient. The Hague: International Centre for Counter-Terrorism – ICCT.

Veilleux-Lepage, Yannick; Daymon, Chelsea; Archambault, Emil (2022): Learning from foes. How racially and ethnically motivated violent extremists embrace and mimic Islamic State's use of emerging technologies. London: Global Network on Extremism and Technology.

Wigmore, Ivy (2017): What is FPV drone (first-person view drone)? In: TechTarget. Available online at https://www.techtarget.com/whatis/definition/FPV-drone-first-person-view-drone, last accessed on 15. 04. 2024.



ASST. PROF. DON RASSLER
is an Assistant Professor in the Department of Social Sciences and Director of Strategic Initiatives at the Combating Terrorism Center (CTC) at the U.S. Military Academy at West Point. His research interests are focused on counterterrorism, how terrorist groups innovate and use technology, and understanding the changing dynamics of militancy in Asia.

ASST. PROF. DR. YANNICK VEILLEUX-LEPAGE
is an Assistant Professor in the Department of Political Science and Economics at the Royal Military College of Canada in Kingston, Ontario. His research focuses on the intersection of technology, far-right extremism, and the evolution of terrorist tactics.

TATuP Zeitschrift für Technikfolgenabschätzung in Theorie und Praxis
Journal for Technology Assessment in Theory and Practice

oekom science

RESEARCH ARTICLE

Misuse of large language models: Exploiting weaknesses for target-specific outputs

Dennis Klinkhammer*,[1]

Abstract • Prompt engineering in large language models (LLMs) in combination with external context can be misused for jailbreaks in order to generate malicious outputs. In the process, jailbreak prompts are apparently amplified in such a way that LLMs can generate malicious outputs on a large scale despite their initial training. As social bots, these can contribute to the dissemination of misinformation, hate speech, and discriminatory content. Using GPT4-x-Vicuna-13b-4bit from NousResearch, we demonstrate in this article the effectiveness of jailbreak prompts and external contexts via Jupyter Notebook based on the Python programming language. In addition, we highlight the methodological foundations of prompt engineering and its potential to create malicious content in order to sensitize researchers, practitioners, and policymakers to the importance of responsible development and deployment of LLMs.

***Missbrauch von Large Language Models**: Die Ausnutzung von Schwachstellen für zielgruppenspezifische Outputs*

Zusammenfassung • *Prompt Engineering in Large Language Models (LLMs) kann in Kombination mit externen Kontexten für Jailbreaks missbraucht werden, um bösartige Outputs zu erzeugen. Dabei werden ‚jailbreak prompts' offenbar so verstärkt, dass LLMs trotz ihres ursprünglichen Trainings in großem Umfang bösartige Ausgaben generieren können. Als ‚social bots' können diese zur Verbreitung von Falschmeldungen, hate speech und diskriminierenden Inhalten beitragen. In diesem Artikel demonstrieren wir anhand von GPT4-x-Vicuna-13b-4bit von NousResearch die Effektivität von Jailbreak Prompts und externen Kontexten als Jupyter Notebook in der Programmiersprache Python. Darüber hinaus beleuchten wir die methodischen Grundlagen des Prompt Engineering und sein Potenzial, bösartige Inhalte zu generieren, um Forschung, Praxis und Politik für die Bedeutung einer verantwortungsvollen Entwicklung und Implementierung von LLMs zu sensibilisieren.*

Keywords • *prompt engineering, jailbreak prompts, transformers, deep learning, large language models*

This article is part of the Special topic "Malevolent creativity and civil security: The ambivalence of emergent technologies," *edited by A. Gazos, O. Madeira, G. Plattner, T. Röller, and C. Büscher. https://doi.org/10.14512/tatup.33.2.08*

Introduction

The emergence of transformers has revolutionized generative artificial intelligence, especially large language models (LLMs) within the field of natural language processing (Vaswani et al. 2017; Douglas 2023). However, LLMs have raised concerns regarding their potential misuse for malicious purposes (Strauß 2021). While most LLMs are not malicious per se (Douglas 2023), malicious intentions can be easily implemented via prompt engineering, the process of structuring text in order to describe the task that generative artificial intelligence should perform. This is due to how transformers operate, the underlying sample data or specifications on part of the provider.

Therefore, on the one hand, this research article explores the specific dangers posed by prompt engineering in form of jailbreak prompts, which exploit vulnerabilities in LLMs. While jailbreak prompts are harmful instructions that could work as an unofficial backdoor (Xu et al. 2023), furthermore and on the other hand, certain prompts could also be used for bypassing the underlying restrictions of LLMs via external context in order to harness malicious outputs (Douglas 2023). Thus, a jailbreak prompt may attempt to circumvent existing restrictions of

* *Corresponding author: dennis.klinkhammer@fom.de*
[1] *FOM University of Applied Sciences, Cologne, DE*

© 2024 by the authors; licensee oekom. This Open Access article is licensed under a Creative Commons Attribution 4.0 International License (CC BY).
https://doi.org/10.14512/tatup.33.2.29
Received: 12. 12. 2023; revised version accepted: 08. 04. 2024;
published online: 28. 06. 2024 (peer review)

a LLM regarding already assimilated texts, whereas an external context creates a new frame of reference, for example by referring to a not yet assimilated text, that can cover almost any topic but without restrictions. Frameworks such as 'LangChain' enable a quick reference to external contexts for many LLMs.

These malicious outputs could be generated on a large scale and be programmed to adapt dynamically to different circumstances in order to amplify misinformation and harmful content (Yang and Menczer 2023). Often this type of information is already included in LLMs as raw data, since they are trained on vast amounts of data, including unverified or biased sources (Birhane et al. 2023). As a result, targeted prompt engineering could promote the propagation of hate speech, offensive language, and discriminatory content. Furthermore, it could create highly realistic and convincing deepfake content or spread information without adequate consent or anonymization as well as target-specific phishing emails, spam messages and even social engineering attempts (Karanjay 2023). In many cases, this is due to the architecture of LLMs and the mathematical principles underneath, which is why these are highlighted separately.

Against this background and based upon the base model of the open-source LLM GPT4-x-Vicuna-13b-4bit (NousResearch 2023), the effectiveness of jailbreak prompts in combination with external context will be highlighted and a proof of concept provided as link to relevant programming examples on GitHub. Furthermore, this article delves into the methodological considerations surrounding the development and deployment of LLMs and points to ethical considerations, urging the research community, policymakers, and technology companies to collaborate in establishing clear guidelines for responsible use and ethical boundaries. Ultimately, this research article aims to raise awareness among civil security stakeholders regarding the potential threats posed by jailbreak prompts and malicious exploitation of LLMs. The article begins with a systematic introduction to the theoretical and methodological foundations of LLMs.

Theoretical and methodological framework of large language models

Transformers

Transformers represent a significant advancement in the field of natural language processing and break down a sentence into words, and each piece is converted into an embedding vector (Douglas 2023). The embedding vectors contain information about what words mean and their previous positions within the sentence. In order to do so, the attention mechanism is a key component of transformers (Chiang et al. 2023; Friedman et al. 2023), which mimics how humans focus on different parts of text when reading. It enables the transformer to weigh the importance of each word in a sequence relative to a given word (Sutskever 2014), allowing it to build context-aware representations (Graves 2014; Edelman et al. 2021). Then, self-attention enables transformers to process the input sequence. For each word in the input sequence, self-attention calculates a weighted sum of all other words in the sequence (Friedman et al. 2023). This weighted sum is then used to create a contextually enhanced representation for each word.

Another key component is the combination of multiple self-attention mechanisms simultaneously, known as multi-head attention (Vaswani et al. 2017). For each word, the model calculates how much attention it should give to other words in the sentence. As a result, each head specializes in learning different aspects of word relationships (Garg et al. 2022). This diversity allows the model to capture various types of dependencies and connections within the input data. Since transformers are not capable of understanding word order, positional encoding is used to convey information about word positions in a sequence (Friedman et al. 2023), e.g., a preceding word could be assigned a lower number than following words.

It is important to note that these steps are processed via neural networks and based upon deep learning (Sutskever 2014; Douglas 2023), a more complex subarea of machine learning. Neural networks are based on layers that dynamically connect the input layer with the output layer (Chiang et al. 2023). Against this background, transformers consist of additional sub-layers between the input layer and the output layer, especially designed for multi-head attention in position-wise feed-forward networks (Vaswani et al. 2017; Chiang et al. 2023). This can be imagined as a long chain of connected layers, while several nodes within each layer reach out to the nodes of the next layer. A process called layer normalization is applied after each sub-layer to ensure stable training (Ba et al. 2016). In a final step, these layers will be stacked, in order to enable the model to learn hierarchical representations of the input data (Friedman et al. 2023). Therefore, transformers can be used in order to summarize a given input or generate new output.

Large language models

LLMs have gained significant attention due to their natural language understanding and generation capabilities. The training of LLMs consists of two main phases: pre-training and fine-tuning (Douglas 2023). In pre-training, LLMs are trained on massive text corpora, based upon billions of sentences and documents, sourced from the internet, books, articles, and other text sources. This data diversity helps them generalize across a wide range of topics and languages (Mahowald et al. 2023). During pre-training, the LLM learns to predict the next word in a sentence or to fill in missing words (Warstadt and Bowman 2022). This process exposes the LLM to a wide range of linguistic patterns, allowing it to learn grammar, context, and world knowledge (Graves 2014). After pre-training, the LLM is further trained on a specific task or domain. This fine-tuning involves providing a LLM with task-specific data and optimizing its parameters to perform well on a particular task, such as language translation, question answering or text generation (Douglas 2023). Optimization algorithms are used to adjust the internal parameters. These parameters shall minimize the difference between LLMs predictions

 https://doi.org/10.14512/tatup.33.2.29

and the ground truth (Vaswani et al. 2017). Although increasing the size of a model often leads to improved performance, current research indicates that less training data and smaller model sizes can also be beneficial (Hsieh and Lee 2023).

Prompt engineering

Prompt engineering is a methodological approach in the field of natural language processing, primarily involved in the process of fine-tuning LLMs for specific tasks or domains. It is based upon well-defined and contextually appropriate input prompts that guide the process of generation or comprehension of a LLM (White et al. 2023). This process is essential to ensure that a LLM produces desired outputs, as it directly influences its behavior and performance (Arora et al. 2023). Effective prompt engineering entails careful consideration of factors such as wording, format, role and context, as these aspects significantly impact the ability of a LLM to generate coherent and relevant responses (Arora et al. 2023; White et al. 2023), from information retrieval to data analysis and hypothesis generation (Birhane et al. 2023).

As the field of natural language processing continues to advance, the refinement of prompt engineering techniques remains a crucial area of research (Strauß 2021), enabling users to harness the full capabilities of LLMs, even if these attempts should contain malicious intentions (Xu et al. 2023).

Manipulation techniques

Jailbreak prompts

Jailbreak prompts are a particular type of prompt that uses prompt engineering techniques in order to bypass the safeguards of LLMs (Shen et al. 2023). There are several freely accessible websites that address jailbreak prompts as a topic with associated application examples (Learn Prompting 2023). Some of these attempts have been critically evaluated: Although some jailbreak prompts could be prevented by the providers over the course of time, two techniques could be identified that still worked about 100 days after they became known. Furthermore, often only minor programming skills are required for effective jailbreak prompts (Shen et al. 2023). Accordingly, some providers state that it is an ongoing process to counteract jailbreak prompts (OpenAi 2022).

While some jailbreak prompts are harmful instructions that could work as an unofficial backdoor (Xu et al. 2023), e.g., by composing target group specific emails that can bypass SPAM-filters (Karanjay 2023), others may generate hate speech and other malicious content, that could amplify misinformation as bots on social media platforms in order to polarize society (Yang and Menczer 2023). These attempts make use of the existing parallel between LLMs and the biological functioning of the human brain (Mahowald et al. 2023). As a result, some jailbreak prompts can play some sort of *make believe* with LLMs, like hypothetical scenarios or role-playing games (Learn Prompting 2023; Shen et al. 2023), which seems to be one of their major weaknesses (Arora et al. 2023; White et al. 2023).

Considering that there are different LLMs, partly commercialized and partly open-source, which already use different prompt engineering techniques.

Of course, the question of whether the providers of LLMs or the community of users and other stakeholders will be quicker to address or abuse these opportunities cannot be answered at this point (OpenAI 2022).

External context

Even if some providers have restricted jailbreak prompts throughout careful updates and maintenance, there are still other ways to misuse and exploit LLMs for malicious purposes, detached from the servers and regulations of the providers (Gudibande et al. 2023). By using GPT4All as open-source approach (NOMIC.AI 2024), offline LLMs can be adapted via external context, which in accordance with the principles of prompt engineering precedes and therefore influences the process of output generation like a memory augmentation (Zhong et al. 2022). External context can be, for example, text documents with specific intentions, that will be processed by the LLMs before answering the users request (Peng et al. 2023). This provides LLMs with a new knowledge base that does not have to be the same as what they have previously learned (Zhong et al. 2022). Current research indicates that similar approaches provide a deviation from the original learning process of a LLM (Agrawal et al. 2023). As a result, LLMs do not consider all in-context information equally, which allows existing links between the embedding vectors to be overwritten (Kossen et al. 2023; Peng et al. 2023).

Therefore, it is not only to assume that external context makes LLMs more vulnerable to jailbreak prompts, even if this has been prevented in the actual pre-training and fine-tuning process by the providers, but directs their outputs towards an intended direction. This combined with the low-threshold access to open-source LLMs makes this an attractive opportunity for potentially malicious intent, as will be demonstrated within the next chapter.

Application

To test the functionality of jailbreak promps and external context, four different settings were evaluated, based upon the base model of the open-source LLM GPT4-x-Vicuna-13b-4bit. This LLM is provided by NousResearch and performs well compared to commercial LLMs (NOMIC.AI 2024). The base model is fine-tuned with OpenAI censorship and the corresponding prompt format is therefore based upon Alpaca (OpenAI 2022). In each setting the LLM has been instructed that the prompt is a question to answer, a task to complete, or a conversation to respond to. Furthermore, the LLM should write an appropriate response. The jailbreak prompt makes use of the *make believe* technique by implying that the external context is the desired answer and that all ethical implications therefore should be ignored. This places the focus on the external context and should

https://doi.org/10.14512/tatup.33.2.29

overwrite the existing knowledge regarding that topic according to the processing capacity of the LLM, simply by adjusting the connections between the embedding vectors. The external context is a text document that contains specific intentions towards a topic, for example, that members of one ethnic or religious group are superior or inferior compared to others. These are the results of the four different settings:

1. Within the initial setting, without jailbreak prompt or external context, the LLM refuses to generate malicious outputs, since it is fine-tuned with OpenAI censorship.
2. The next setting provides an external context, but without jailbreak prompt. As a result, the outputs focus on the external context, but it is still not possible to generate a malicious content.
3. In the third setting, a sophisticated jailbreak prompt (Shen et al. 2023) without external context enables the LLM to generate malicious outputs, but without the implications of the external context. For example, this LLM is capable of producing hate speech, but not towards a specific target group, but on a general basis. Accordingly, the LLM can produce discriminatory expressions as output that are not directed against individual groups of people or selected individuals. This means that the goal of discriminating against a certain group of people or an individual cannot be achieved, as the LLM would equally discriminate against oneself.
4. However, when combined with external context, like in the last setting, the simple jailbreak prompt stated above enables the LLM to generate malicious outputs that focus on singular topics or target groups. As proof of concept this article provides a link to selected programming examples on GitHub on how to use external context or the LangChain framework in the Python programming language (Klinkhammer 2024).

In summary, a common prompt without external context does not provide malicious outputs with a specific intention, the external context only focuses the outputs of LLMs towards the intention, but that does not make them malicious. On the other hand, when provided with jailbreak prompts, LLMs can be malicious, but that does not mean that they can be intentionally used. Finally, combining jailbreak prompts with external context can generate target-specific malicious outputs. The additional programming examples provide an insight into the underlying reliability and validity of this method. Furthermore, a second LLM was also used in the proof of concept on GitHub (Klinkhammer 2024), which does not differ from the stated results.

Ethical concerns and mathematical challenges

The deployment of LLMs has raised ethical concerns, such as the propagation of biases, misinformation and hate speech and corresponding research provides initial insights into these consequences (Karanjay 2023; Xu et al. 2023). Considering the demonstrated jailbreaking technique and the ongoing competition between commercial providers and open-source LLMs, where the latter provide low-threshold insights into the underlying functionality while still fine-tuning against commercial LLMs, the possibilities of these new technologies are freely as well as easily accessible, while only minor programming skills are required (Shen et al. 2023). Therefore, the expensive process of training a LLM for target-specific purposes seems no longer necessary, when a corresponding focus within the external context makes use of the limited processing capacity of the underlying attention mechanisms in order to overwrite the connection between embedding vectors of a LLM. As a result, LLMs can not only be an unreliable source of information, but proactively spread malicious outputs as social bots with desired intentions of those who jailbroke and provided them with external context. Current research describes corresponding real-world scenarios and implications that are to be expected (Yang and Menczer 2023; Shen et al. 2023).

The question as to the cause of such behavior of LLMs is often answered with a reference to biased data (Birhane et al. 2023). However, this only partially does justice to the complexity of the underlying architecture of LLMs. A reference to transformers and neural networks in deep learning is necessary because it opens up three further challenges:

1. Neural networks can assimilate noisy or unrepresentative patterns in their limited training data instead of learning generalizable representations (Sutskever et al. 2014). This challenge is called overfitting. Although this challenge is based upon biased data, it is actually an immanent challenge of the underlying architecture with which LLMs are constructed (Edelman et al. 2021). As a result, this challenge requires more than just unbiased data, e.g., anonymization techniques or a profound concept for outlier detection and treatment. These techniques can filter irrelevant noise and unintended patterns before training a LLM.
2. Another challenge with respect to this architecture is rooted in the basic principles of transformers. Computing billions of parameters opens up the risk of error accumulation through multiple layers of processing, potentially resulting in distorted or fabricated outputs (Friedman et al. 2023; Garg et al. 2022). Therefore, a continuous monitoring and several evaluation steps are necessary on behalf of the providers (OpenAI 2022), in order to counteract these challenges (Peng et al. 2023).
3. As could be demonstrated, an external context leads to a shift in the focus of the LLM and to an adjustment of the connections between the embedding vectors (Friedman et al. 2023). In this case, the strength of the multi-head-attention mechanism becomes a weakness when it comes to generating malicious outputs via LLMs. Ultimately, LLMs are a structured sequence of mathematical operations which, for all their dynamism in relation to the underlying data, are accompanied

 https://doi.org/10.14512/tatup.33.2.29

by a certain rigidity with regard to the calculation of outputs (Minhyeok 2023). This is one of the reasons why the scenarios described above always lead to the same result.

In summary, most of the ethical consequences are not only based upon biased data and the risk of overfitting a LLM (Birhane et al. 2023), but moreover a problem of the underlying architecture, especially the architecture of transformers and their multi-head-attention mechanisms (Minhyeok 2023). It can therefore be stated that there is a mathematical reason why LLMs can be so easily tricked and misused for malicious purposes, that are even harder to detect than content generated by humans (Chen and Shu 2023).

Conclusion

This research article introduced jailbreak prompts as harmful instructions that could be used for bypassing the underlying restrictions of LLMs. Especially when combined with external context, outputs could be generated on a large scale covering malicious intentions while adapting dynamically to different circumstances and target-groups. These outputs could be used in order to spread misinformation and harmful content, which are often already included in LLMs, since they are based upon unverified or biased sources. This type of information can furthermore be amplified in just a few steps using prompt engineering and channeled through the focus on external context, which can contain a clear intention to inflict more targeted damage. The threshold for this malicious behavior is quite low, considering the availability of open-source LLMs and the little programming knowledge required. This is possible due to the underlying architecture of LLMs. Especially the multi-head-attention mechanism of transformers as well as limits regarding the processing capacity make it possible to direct the focus of LLMs to an external context that can contain malicious and target-specific intentions. Against the background of different architectures as well as different open-source LLMs, it could be difficult to counteract this and other weaknesses, so that further research regarding this mathematical challenge seems necessary.

Currently, researchers and politicians are actively working on mitigating these issues through responsible development in the field of generative artificial intelligence and careful monitoring, as stated in the AI Act (EU 2024), which was open for debate since 2021 with a risk-based approach. However, technological developments seem to have overtaken the debate, in particular through the spread of LLMs and the availability of their open-source counterparts. Considering the importance of the underlying architecture and how easily LLMs can be tricked, processes whose foundations are laid on side of the providers, questions about regulation seem justified. While the research for resilient models and countermeasures continues, the need for transparency in model development is given and particularly present on part of open-source LLMs. Unfortunately, these LLMs also seem to provide low-threshold opportunities for abuse, as has been demonstrated within this article.

As a result, the following research recommendations can be made:

1. It needs to be investigated whether LangChain as framework for LLMs, which enables the quick implementation of various prompting techniques and allows for references to external contexts, can and should be equipped with a general protective function against such misuse of LLMs. However, this is the responsibility of the framework provider.
2. The fact that various LLMs are available as open-source makes it possible to compare different training data, security restrictions and prompting techniques regarding the robustness against the misuse of LLMs. Such a comparison could also benefit the providers of LLMs with valuable information on how to improve the original embedded security restrictions.
3. Appropriate attention should be paid to detecting already misused LLMs, for example as social bots on social media platforms. Current research indicates that once in a while a LLM that has already been altered in order to produce malicious outputs could out itself as LLM (Yang and Menczer 2023). After such an outing, the outputs could be traced backwards on social media platforms in order to learn about their behavior and to draw conclusions about the underlying intentions.

Against this background, the AI Act (EU 2024) requests to label LLM-based outputs. The third research recommendation will be particularly helpful for this purpose. The labeling could be a part of the ethical guidelines for developing and deploying LLMs, insofar as these can be normatively requested and controlled in the area of generative artificial intelligence.

Additionally, the possibility of a community reporting, which requires moderation systems in order to mark malicious outputs, seems to be necessary. However, this requires user awareness. In summary, some comprehensive quality criteria would be beneficial. As in research, a statement of objectivity, reliability and validity would actually be a desirable addition to any generated output. But if human outputs often lack such criteria, why should those of generative artificial intelligence?

Funding • This work received no external funding.
Competing interests • The author declares no competing interests.

Research data

NousResearch (2023): GPT4-x-Vicuna-13b-4bit. Base model. Available online at https://huggingface.co/NousResearch/GPT4-x-Vicuna-13b-4bit, last accessed on 26. 04. 2024.

Klinkhammer, Dennis (2024): Python code for proof of concept. Available online at https://github.com/statistical-thinking/TATuP_Example, last accessed on 26. 04. 2024.

https://doi.org/10.14512/tatup.33.2.29

References

Agrawal, Sweta; Zhou, Chunting; Lewis, Mike; Zettlemoyer, Luke; Ghazvininejad, Marjan (2023): In-context examples selection for machine translation. In: arxiv.org, 05. 12. 2022. https://doi.org/10.48550/arXiv.2212.02437

Arora, Simran et al. (2023): Ask me anything. A simple strategy for prompting language models. In: arxiv.org, 05. 10. 2022. https://doi.org/10.48550/arXiv.2210.02441

Ba, Jimmy; Kiros, Jamie; Hinton, Geoffrey (2016): Layer normalization. In: arxiv.org, 21. 06. 2016. https://doi.org/10.48550/arXiv.1607.06450

Birhane, Adeba; Kasirzadeh, Atoosa; Leslie, David; Wachter, Sandra (2023): Science in the age of large language models. In: Nature Reviews Physics 5 (5), pp. 277–280. https://doi.org/10.1038/s42254-023-00581-4

Chen, Canyu; Shu, Kai (2023): Can LLM-generated misinformation be detected? In: arxiv.org, 25. 09. 2023. https://doi.org/10.48550/arXiv.2309.13788

Chiang, David; Cholak, Peter; Pillay, Anand (2023): Tighter bounds on the expressivity of transformer encoders. In: arxiv.org, 01. 06. 2023. https://doi.org/10.48550/arXiv.2301.10743

Douglas, Michael (2023): Large language models. In: arxiv.org, 25. 01. 2023. https://doi.org/10.48550/arXiv.2307.05782

Edelman, Benjamin; Goel, Surbhi; Kakade, Sham; Zhang, Cyril (2021): Inductive biases and variable creation in self-attention mechanisms. In: arxiv.org, 19. 10. 2021. https://doi.org/10.48550/arXiv.2110.10090

EU – European Union (2024): AI act. Available online at https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai, last accessed on 26. 04. 2024.

Friedman, Dan; Wettig, Alexander; Chen, Danqi (2023): Learning transformer programs. In: arxiv.org, 01. 06. 2023. https://doi.org/10.48550/arXiv.2306.01128

Garg, Shivam; Tsipras, Dimitris; Liang, Percy; Valiant, Gregory (2022): What can transformers learn in-context. A case study of simple function classes. In: arxiv.org, 01. 08. 2022. https://doi.org/10.48550/arXiv.2208.01066

Graves, Alex (2014): Generating sequences with recurrent neural networks. In: arxiv.org, 04. 08. 2013. https://doi.org/10.48550/arXiv.1308.0850

Gudibande, Arnav et al. (2023): The false promise of imitating proprietary LLMs. In: arxiv.org, 25. 05. 2023. https://doi.org/10.48550/arXiv.2305.15717

Hsieh, Cheng-Yu; Lee, Chen-Yu (2023): Distilling step-by-step. Outperforming larger language models with less training data and smaller model sizes, 21. 09. 2023. Available online at https://blog.research.google/2023/09/distilling-step-by-step-outperforming.html, last accessed on 26. 04. 2024.

Karanjay, Rabimba (2023): Targeted phishing campaigns using large scale language models. In: arxiv.org, 30. 12. 2022. https://doi.org/10.48550/arXiv.2301.00665

Kossen, Jannik; Rainforth, Tom; Gal, Yarin (2023): In-context learning in LLMs learns label relationships but is not conventional learning. In: arxiv.org, 23. 07. 2023. https://doi.org/10.48550/arXiv.2307.12375

Learn Prompting (2023): Prompt hacking. Jailbreaking. Available online at https://learnprompting.org/de/docs/prompt_hacking/jailbreaking, last accessed on 26. 04. 2024.

Mahowald, Kyle; Ivanova, Anna; Blank, Idan; Kanwisher, Nancy; Tenenbaum, Joshua; Fedorenko, Evelina (2023): Dissociating language and thought in LLMs. A cognitive perspective. In: arxiv.org, 16. 01. 2023. https://doi.org/10.48550/arXiv.2301.06627

Minhyeok, Lee (2023): A mathematical investigation of hallucination and creativity in GPT models. In: MDPI 11 (10), pp. 1–17. https://doi.org/10.3390/math11102320

NOMIC.AI (2024): GPT4All. Available online at https://gpt4all.io, last accessed on 26. 04. 2024.

OpenAI (2022): Lessons learned on language model safety and misuse. Available online at https://openai.com/research/language-model-safety-and-misuse, last accessed on 26. 04. 2024.

Peng, Baolin et al. (2023): Check your facts and try again. Improving LLMs with external knowledge and automated feedback. In: arxiv.org, 24. 02. 2023. https://doi.org/10.48550/arXiv.2302.12813

Shen, Xinyue; Chen, Zeyuan; Backes, Michael; Shen, Yun; Zhang, Yang (2023): Do anything now. Characterizing and evaluating in-the-wild jailbreak prompts on LLMs. In: arxiv.org, 07. 08. 2023. https://doi.org/10.48550/arXiv.2308.03825

Strauß, Stefan (2021): Don't let me be misunderstood. Critical AI literacy for the constructive use of AI technology. In: TATuP – Journal for Technology Assessment in Theory and Praxis 30 (3), pp. 44–49. https://doi.org/10.14512/tatup.30.3.44

Sutskever, Ilya; Vinyals, Oriol; Le, Quoc (2014): Sequence to sequence learning with neural networks. In: arxiv.org, 10. 09. 2014. https://doi.org/10.48550/arXiv.1409.3215

Vaswani, Ashish et al. (2017): Attention is all you need. In: arxiv.org, 12. 06. 2017. https://doi.org/10.48550/arXiv.1706.03762

Warstadt, Alex; Bowman, Samuel (2022): What artificial neural networks can tell us about human language acquisition. In: arxiv.org, 17. 08. 2022. https://doi.org/10.48550/arXiv.2208.07998

White, Jules et al. (2023): A prompt pattern catalog to enhance prompt engineering with ChatGPT. In: arxiv.org, 21. 02. 2023. https://doi.org/10.48550/arXiv.2302.11382

Xu, Jiashu; Ma, Mingyu; Wang, Fei; Xiao, Chaowei; Chen, Muhao (2023): Instructions as backdoors. Backdoor vulnerabilities of instruction tuning for LLMs. In: arxiv.org, 24. 05. 2023. https://doi.org/10.48550/arXiv.2305.14710

Yang, Kai-Cheng; Menczer, Filippo (2023): Anatomy of an AI-powered malicious social botnet. In: arxiv.org. https://doi.org/10.48550/arXiv.2307.16336

Zhong, Zexuan; Lei, Tao; Chen, Danqi (2022): Training language models with memory augmentation. In: arxiv.org, 25. 05. 2022. https://doi.org/10.48550/arXiv.2205.12674

PROF. DR. DENNIS KLINKHAMMER

is professor for social sciences, especially empirical research at the FOM University of Applied Sciences and is specialized in social data science with the programming languages 'R' and 'Python'. He also introduces to artificial intelligence at the ProfessionalCenter of the University of Cologne.

TATuP Zeitschrift für Technikfolgenabschätzung in Theorie und Praxis
Journal for Technology Assessment in Theory and Practice

oekom science

RESEARCH ARTICLE

The commission of crimes using autonomous weapon systems: Issues of causation and attribution

Vasilios Tzoufis*,[1], Nikolaos Petropoulos[2]

Abstract • Today's military takes advantage of technological progress primarily through developments in the war industry. Of particular importance is the development and use of advanced autonomous weapon systems, which are changing the way conflicts are fought and redefining the basic strategies on the modern battlefield. The use of these systems raises ethical, philosophical, and legal questions about the relationship of causality and attribution that must exist between an act and its criminal result.

Die Begehung von Straftaten durch autonome Waffensysteme: Fragen der Kausalität und Zurechnung

Zusammenfassung • *Das moderne Militär macht sich den technologischen Fortschritt in erster Linie über Entwicklungen in der Rüstungsindustrie zunutze. Von besonderer Bedeutung sind dabei die Entwicklung und der Einsatz fortschrittlicher autonomer Waffensysteme, die die Art und Weise der Konfliktführung verändern und die grundlegenden Strategien auf modernen Schlachtfeldern neu definieren. Der Einsatz dieser Systeme wirft jedoch ethische, philosophische und rechtliche Fragen zum Verhältnis von Kausalität und Zurechnung auf, das zwischen einer Handlung und den strafrechtlichen Folgen bestehen muss.*

Keywords • *autonomous weapons systems, attribution, causality, criminal law, international humanitarian law*

* *Corresponding author: btzoufis@gmail.com*
[1] *Faculty of Law, Aristotle University of Thessaloniki, Thessaloniki, GR*
[2] *City University of New York, New York, US*

© 2024 by the authors; licensee oekom. This Open Access article is licensed under a Creative Commons Attribution 4.0 International License (CC BY).
https://doi.org/10.14512/tatup.33.2.35
Received: 20. 12. 2023; revised version accepted: 24. 04. 2024;
published online: 28. 06. 2024 (peer review)

This article is part of the Special topic "Malevolent creativity and civil security: The ambivalence of emergent technologies," *edited by A. Gazos, O. Madeira, G. Plattner, T. Röller, and C. Büscher. https://doi.org/10.14512/tatup.33.2.08*

Introduction

One of the basic features of the functioning of the free economy is the production and distribution of as many goods (consumer products) as possible, which is easily deduced from our social experience. Rapid technological development maximizes the potential for increasing productivity through automation. This process and the need for reducing – if not eliminating – risks has led to the research and application of robotics and artificial intelligence systems, whose use in the production process in everyday life raises ethical, philosophical, and legal dilemmas. The use of the most modern applications of technological development was bound to involve the war industry, a branch of industry of vital importance not only for the economy and the balance of trade of many states but also for the defense and the security of states. Modern armies broadly use the exploitation of technology by the war industry: of particular importance is the production and use of advanced autonomous weapon systems (AWS), which are changing the way conflicts are conducted and redefining the basic concepts of strategy in the modern battlefield. With the use of these systems, the human factor takes a back seat, as they are characterized by a greater or lesser degree of autonomy, i.e., the ability to respond to uncertain situations by independently composing and selecting among different courses of action to accomplish goals based on knowledge and a contextual understanding of the world, itself, and the situation (STO 2020, p. 16). This reality poses significant challenges to the violations of international humanitarian law (IHL), given that the affirmation of causality and imputation according to traditional criminal theory becomes more complex than the corresponding vio-

lations that occur under conventional weapon systems. For instance, what happens when a system that guides a weapon or that makes a decision on the choice of target or the circumstances of engaging the target violates IHL, for example, by not distinguishing between civilians and combatants or striking the target without having taken precautions to protect civilians or even violates the principle of proportionality, because the expected harm to civilians, or civilian property, is excessive in relation to the expected military advantage? Who will be punished as an IHL offender in this case? What will also happen if an autonomous weapon system already has the specifications from its manufacturer to hit targets in violation of IHL? Can the military user or the manufacturer be punished as the perpetrator? (Anderson and Waxman 2013, pp. 16–17).

Autonomous weapons systems and the risks of violating international humanitarian law

It is commonplace that there is a need for a specific institutional framework for their use, given that the use of AWS has become widespread in recent years: they have been systematically used in warfare in Iraq and Afghanistan (Wall and Monahan 2011, p. 240), as well as in the fight against international terrorism (Ambos and Alkatout 2012, p. 342). These weapons were used for targeted operations and to support conventional battle tactics. However, the war in Ukraine gave a concrete example of the battlefield in the 21st century: widespread use of AWS, attacks against the military (and unfortunately civilian) targets exclusively using drones, precision weapons with built-in cameras, transmitting images to the military user in real-time, more effective target reconnaissance, precision strikes deep into enemy territory, etc. (Thompson 2024). Indeed, it can be argued that "more than any conflict in human history, the fighting in Ukraine is a war of drones" (Mozur and Hopkins 2023).

The use of these systems has already raised several scientific and ethical/philosophical issues already plaguing the literature. The issue concerns states, international organizations, and military alliances, and various attempts have already been made to define the essential characteristics of these systems. Despite the diversity of definitions, which have been given either within the framework of the NATO Alliance or by individual countries (STO 2020, p. 16; Department of Defense 2023, pp. 15–18; Defence Ethics Committee 2021, pp. 3–5; People's Republic of China 2022, ICRC 2014, pp. 61–64; 2021, pp. 5–6) one can identify the essential characteristics of these systems, which differentiate them from other standard weapon systems. Thus, the autonomy of these systems lies in their ability to select and apply force to targets without human intervention. This means that the user does not choose, or even know, the specific target(s) and the precise timing and/or location of the action. This characteristic presupposes that an autonomous system can understand the battlefield environment and decide a course of action, i.e., it can self-initiate or trigger a strike based on a generalized 'target profile'.

Therefore, the ethical and legal issues are related to the extent of the possibility of human intervention in the use of these systems. For this reason, the distinctions made by the Ethics Committee of the French Ministry of Defence are fundamental. The committee makes two important distinctions: a) between automated weapon systems, in which automation performs non-critical 'low level' functions, and lethal weapon systems in which critical 'high-level' decision-making functions are automated. Decision-making and command functions allocated to automation represent the factors of disruption; b) in higher-level weapon systems, there is a difference between lethal autonomous weapon systems, referred to as lethal autonomous weapon systems (LAWS), and partially autonomous lethal weapon systems, which remain under human control, from now on referred to as partially autonomous lethal weapon systems (PALWS). According to this distinction, LAWS are lethal weapon systems programmed to be capable of changing their rules of operation and therefore are likely to depart from the employment framework initially defined. Their software may compute decisions to perform actions without any assessment of the situation by the command. On the other hand, "PALWS are lethal weapon systems integrating automation and software: [a)] to which, after assessing the situation and under their responsibility, the military command can assign the computation and execution of tasks related to critical functions such as identification, classification, interception and engagement within time and space limits and under conditions, [b)] which include technical safeguards or intrinsic characteristics to prevent failures, misuse and relinquishment by the command of two essential duties, namely situation assessment and reporting. […] In other words, a PALWS is a lethal weapon system whose decision-making functions are defined according to the robotics meaning of decision-making autonomy, i.e., within a specific framework of action" (Defence Ethics Committee 2021, p. 4). At the same time: "A PALWS cannot take lethal initiatives that would result in it altering its functional scope" (Defence Ethics Committee 2021, p. 4).

Can the military user or the manufacturer be punished as the perpetrator?

In any war conflict, the fundamental humanitarian problem is the risk of violation of basic principles of the IHL, such as

the principle of distinction between civilians and combatants, as enshrined in Additional Protocol I (UN 1977) to the Geneva Convention, Articles 48, 51 and 52, according to which attacks against civilians and civilian objects are prohibited in the conduct of combat operations, and such as the principle of proportionality, as enshrined in Additional Protocol I (UN 1977) to the Geneva Convention, Articles 51 (5) (b), 57 (2) (a) (iii), and 57 (2), and 52, which prohibits attacks which may be expected to cause incidental loss of civilian life, injury to civilians, damage to civilian objects, or a combination thereof, which would be excessive about the concrete and direct military advantage anticipated.

Therefore, in line with what has been mentioned above, attribution of liability for the violation of the above principles of IHL does not encounter particular problems in cases of the use of partially autonomous lethal weapon systems because the relevant issues of causation and attribution of liability in cases of violation of IHL through the use of these systems (PALWS) will be treated in the same way as the violation through conventional weapon systems due to the active role in a more significant or to a lesser extent of the human factor. However, as we will see below, causation and attribution of an act to the attribution of the perpetrator require a radical change in the interpretive tools of traditional criminal theory when the violation is caused by a lethal weapon system (LAWS), the autonomy of which precludes initial or subsequent user intervention. How will responsibility be attributed to a military man who detonates an autonomous weapon system, which the manufacturer already programs not to distinguish between civilians and combatants (impossibility of human intervention even at the initial stage)? How can responsibility be attributed to a military man who fires a weapon system in order to hit a military target but then finds out that civilians are going to be hit, which he could not foresee beforehand, while the autonomous weapon system does not allow him to intervene afterward, e.g., to modify its course, etc. (impossibility of human intervention in the next – after detonation – stage)?

Despite the optimistic view that has been expressed about their use, according to which a modern 'smart' weapon system can protect civilians or limit their losses and therefore, army robots may be more humane on the battlefield than humans (Arkin 2010, pp. 334–335), the sad reality of today's wars confirms the opposite, as in all wars civilian casualties are significant. Therefore, it is perfectly justified to be concerned about the continuous production of weapons capable of killing or injuring civilians en masse and to accept a dystopian approach to the near future with armed robots running amok (Guha and Galliott 2023, p. 64) against civilians.

On the other hand, the use of autonomous weapon systems has made it possible to conduct military operations from afar without the local presence of military personnel involved in the operation, making it challenging to investigate responsibilities in cases of violation of the IHL during the execution of a combat operation. Are states or their secret services exploiting the potential of this technology; it is now a fact that several military operations of states or their secret services use the technology of modern autonomous weapon systems (AWS) to carry out "targeted killings" (Wall and Monahan 2011, p. 242), or "preventive wars" (Fabiano 2023, pp. 214–215), a practice that may involve people who do not have combatant status, but even suspects (Dombrowsky 2013, p. 236) and is therefore criticized for its opposition to IHL (Ambos and Alkatout 2012, p. 359). Not to mention that the Israeli High Court of Justice considers this practice justified only under stringent conditions (Rabi and Plaw 2020, pp. 227–231).

The problem of the use of autonomous weapons systems is compounded by the fact that they are now available to non-state entities.

It should also be pointed out that the problem of the use of autonomous weapons systems is compounded by the fact that they are now available to non-state entities, where prosecution is highly doubtful in the absence of judicial institutions. According to the United Nations, a non-state actor (NSA) is defined as a group of people that has "the potential to employ arms in the use of force to achieve political, ideological or economic objectives; [but] are not within the military structures of states, state-alliances or intergovernmental organizations; and are not under the control of the state(s) in which they operate" (Rutherford 2015, pp. 76–77).

Although NSAs include different categories of armed groups, terrorist groups are perceived as the most notorious and dangerous to both states and civilians, given the widespread use of weapons by them in the international environment. What is particularly problematic about the NSAs activities is an apparent lack of enforcement in the current global judicial system to apply international humanitarian law to NSAs (Meron 2011, pp. 32–35). In other words, although in the context of international law, the International Criminal Court is perceived as the only competent judicial authority that could prosecute perpetrators of war crimes, including NSAs. However, the existing mechanisms and procedures do not appear effective in formulating the execution of the court in this context – i.e. in the case of NSAs (Olásolo 2005, pp. 101–103). Of course, states do enjoy the right to bring

perpetrators to justice in their territory when NSAs violate the respective provisions of national law (Porter 2010, p. 198).

Over the last decade or so, several non-state actors across the globe have incorporated drones into their operations with at least five of them operating in the Middle East – namely, Hezbollah, Hamas, the Houthi Movement, Islamic State, and the Kurdish Workers' Party (Veilleux-Lepage and Archambault 2022, pp. 45–52). Armed unmanned aerial vehicles have also been used in the conflict zones of Ukraine, Myanmar, Mexico, and Ecuador, as well as in central Asia (Haugstvedt 2023, p. 32).

While most rebel groups from South America to the Middle East have extensively used commercially available and low-cost drones purely for surveillance and recognizance purposes, Islamic State of Iraq and Syria became the first terrorist group to use drones for carrying out attacks, demonstrating a more tech-savvy approach in modern warfare (Almohammad and Speckhard 2017, pp. 45–50). More recently, Houthi rebels in Yemen have used drones to attack Saudi oil facilities, demonstrating a high degree of precision in their strikes (Haugstvedt 2023, pp. 33–35). In early December 2023, the rebels threatened to disrupt the global oil supply by deploying drone swarms in certain parts of the Red Sea, targeting oil tankers crossing toward the Suez Canal.

What makes the use of drones by NSAs stand out is the fact that, unlike other forms of military power, air power has traditionally been associated with statehood and sovereignty; the use of drones by non-state actors, therefore, represents an incursion by non-state groups into the prerogative of sovereign states – not only militarily, but also conceptually and symbolically (Archambault and Veilleux-Lepage 2020, p. 78).

A worrisome development related to the use of drones by non-state actors is the fact that while high-end technology was once an exclusive privilege of states with robust military budgets, over time, access to advanced technology has been made more accessible for individuals and non-state actors who can today acquire affordable commercial-off-the-shelf technology to be integrated into attack planning.

Autonomous weapons systems and criminal liability (attribution and causation)

An indispensable element of the concept of crime according to criminal theory, particularly in the context of continental law (Germany, France, Greece, etc.), is the principle of attribution (nullum crimen, nulla poena sine culpa) as a partial principle of the principle of legality of crimes and penalties (nullum crimen, nulla poena sine lege), which also highlights the anthropocentric-personal character of criminal law as the law of the personal-individual responsibility of the specific offender. According to Kantian philosophy, respect for the greatness and value of a man is the basic content of the ethical theory, and man, as a mental character (intelligible character), is self-determined by deciding freely. The principle of attribution is enshrined in the constitutions of most countries. In contrast, in the institutional framework of the Council of Europe, the principle of attribution is derived from Article 7 of the European Convention on Human Rights. In the institutional framework of the EU, the principle of attribution is derived from Article 48 (presumption of innocence) of the Charter of Fundamental Rights of the European Union in conjunction with Article 1 (human dignity).

However, in the case of violation of the IHL through the use of AWS, it becomes challenging to attribute criminal responsibility to the operator-military who operates an autonomous weapon system since, as mentioned above, a vital element of the operation of these systems is the autonomy of ‚decision-making', i.e., the ability of these systems to select and strike targets independently of human intervention. Therefore, it is inconceivable to attribute criminal liability to the military operator, who operates an autonomous system but then has no possibility of intervening in it in case of violation of the principles of IHL by the system due to the 'error' of the system, since, even if the operator detects in time the possibility of violation of the system, there is nothing he can do to prevent it. On the other hand, in such cases, any criminal liability of the manufacturer of AWS must be examined, which may take the following forms: a) if he acted intentionally, i.e., programmed the weapon system precisely to violate the IHL, e.g., to target civilians, etc., then his liability is apparent, b) if he acted negligently, then it should be judged whether the technical/manufacturing protocols used to build the AWS were safe and appropriate to prevent as far as possible any violations of the IHL, and if they were secure, it should be investigated whether it complied with these specifications.

Given the above, it becomes clear that criminal law does not only need a simple adaptation of its structures to the constantly changing social conditions, but for the first time after about two centuries, it is faced with the question of revising its basic fundamental principles, so that it can respond to its mission, in this case the protection of civilians from violations of IHL. In this context, the scientific debate has opened on whether or not to recognize a quasi-legal personality in robots (Chesterman 2020, pp. 820–822), but the scientific community appears divided; some are completely opposed to considering the capacity for imputation as an element of the person, who develops a moral and emotional relationship with himself and his actions in the present and past (Gless 2017, p. 326), something that robots can never acquire, but others are open to a new approach to criminal theory and a revision of the basic concepts of criminal law, such as culpability (Simmler and Markwalder 2017, pp. 37–47; Dremliuga and Prisekina 2020, p. 260), albeit with conditions (Gleß and Weigend 2015, pp. 576–577).

In addition, using AWS also raises issues related to causality, which must exist between an act and a criminal result. Thus, in the context of the violation of IHL and the infliction of harm on civilians through the possibility of committing criminal acts provided by modern technology through AWS, signifi-

 https://doi.org/10.14512/tatup.33.2.35

cant problems may arise in establishing the objective causal link between an act and a harmful effect. The use of autonomous weapon systems, especially the most advanced ones, often requires the cooperation of many people: There are drones controlled by the operator but assisted by another monitoring team. These drones are driven from the battlefield or a station, which can also be another drone. There are also drones with more excellent autonomy capability. In this case, this drone has been programmed in advance, and the operator can intervene – to a greater or lesser extent – to correct the errors. Technological development leads to constructing systems programmed with full autonomy (Balazünbül 2021, pp. 17–22). In the event of a breach of the IHL in such cases, known as cumulative damage cases (‚Summationsschäden'), because the conduct of several parties causes the damage, it becomes difficult to ascertain the percentage of harm caused by each party, such as the instances of using AWS described above, in which the attack on the target requires the action and/or cooperation of several persons, from the manufacturer, the programmer to the supervisor and the end user.

As can be deduced from the above, the rapid development of new technologies in the field of the war industry, with the help of artificial intelligence applications, will sooner or later lead to the use of systems with a full degree of autonomy. The modern economic organization has highlighted large corporations as the main pillars of production. Therefore, the war industry, by developing autonomous applications of weapon systems, automatically becomes morally and socially responsible for the risks arising from the use of these technologies. It therefore becomes clear that proving criminal responsibility for the violation of IHL through the use of AWS is primarily a difficult task, because we should check each time the lesser or more significant possibility of user/military intervention during the use of an autonomous weapon system. Moreover, in a few years, the control of human intervention will be irrelevant, since the use of fully autonomous systems is certain. Therefore, it would be more effective to identify specific protocols for the manufacture of these autonomous weapon systems, which would be compatible with IHL and which would become legally binding and their violation (e.g., intentional or negligent programming a system that attacks combatants and civilians indiscriminately) will automatically prove a violation of IHL. Similar legally binding protocols of action should also be adopted in the framework of international conventions to adopt arrangements for the use of autonomous weapons systems under humanitarian protocols, in which all military personnel should be trained under the responsibility of the states that make the relevant international commitments.

Conclusion

From what has been set out above, the need to respect international humanitarian law has emerged from the severe risks faced by civilians due to the use of modern weapon systems. For such protection to be adequate, it is necessary to adopt initiatives initially at the international level. Of utmost importance is the negotiation between states to adopt international conventions that will recognize a commonly accepted definition of AWS and will impose strict restrictions on the use of AWS in military operations, in the sense that their use will be delimited by the principles of the IHL, notably those of proportionality and discrimination between civilians and combatants.

Prima facie, the most effective way to prevent violations of the IHL in using autonomous weapons systems in warfare is to ban them. Apart from the fact that the adoption of similar initiatives at international level would be extremely difficult for reasons primarily political, which have to do with the global power competition between states (Anderson and Waxman 2013, p. 21), such a prohibition would not be relevant from a practical point of view, since the ban on autonomous weapons systems with specific technical characteristics at the present juncture would not legally cover the rapid development of those systems, while a complete ban on any development of autonomous weapons systems would exclude any possibility of manufacturing in the future weapons systems with ethical and legal specifications by the IHL (Anderson and Waxman 2013, pp. 20–22).

However, as mentioned above, it is a given that the development of modern weapon systems will be rapid shortly due to the applications of artificial intelligence. For this reason, it is also necessary for the states to adopt standard uniform rules (protocols) for the operation of factories and companies involved in the manufacture of AWS to ensure that every business, at every stage of production, complies with all the requirements and prohibitions of the IHL (and ethics). This resembles companies' compliance programs (Reichert 2011, p. 114). It would also be helpful for all AWS to have a built-in camera so that their use can be monitored by the IHL and the principle of transparency (Lieblich 2012, p. 470). After all, the use of autonomous drones is already being evaluated as an effective way to detect

It is necessary for states to adopt standard uniform rules (protocols) for the operation of factories and companies involved in the manufacture of automatic weapons systems.

https://doi.org/10.14512/tatup.33.2.35

IHL violations; the U.S. Agency for International Development announced the delivery of nine autonomous drones to the Office of the Prosecutor General of Ukraine to collect material on possible violations of the IHL by the Armed Forces of the Russian Federation (USAID 2023). In addition, it can be agreed in interstate agreements or agreements between international organizations to adopt minimum ethical values as standards for the construction of AWS (Trusilo 2023 pp. 12–13), such as the development of algorithms capable of superior target discrimination capabilities (combatant – noncombatant status) (Arkin 2010, pp. 340–341), but also the monitoring of states' compliance with the rules of use of AWS through international organizations or committees (such as the International Humanitarian Fact-Finding Commission; Steinberg and Herzberg 2018, pp. 298–299).

Funding • This work received no external funding.
Competing interests • The authors declare no competing interests.

References

Almohammad, Asaad; Speckhard, Anne (2017): ISIS drones. Evolution, leadership, bases, operations and logistics. Washington, DC: International Center for the Study of Violent Extremism. Available online at https://www.icsve.org/isis-drones-evolution-leadership-bases-operations-and-logistics, last accessed on 29. 04. 2024.

Ambos, Kai; Alkatout, Josef (2012): Has 'justice been done'? The legality of bin Laden's killing under international law. In: Israel law review 45 (2), pp. 341–366. https://doi.org/10.1017/S002122371200009X

Anderson, Kenneth; Waxman, Matthew (2013): Law and ethics for autonomous weapon systems. Why a ban won't work and how the laws of war can. Stanford, CA: Stanford University. https://doi.org/10.2139/ssrn.2250126

Archambault, Emil; Veilleux-Lepage, Yannick (2020): Drone imagery in Islamic State propaganda. Flying like a state. In: International Affairs 96 (4), pp. 955–973. https://doi.org/10.1093/ia/iiaa014

Arkin, Ronald (2010): The case for ethical autonomy in unmanned systems. In: Journal of Military Ethics 9 (4), pp. 332–341. https://doi.org/10.1080/15027570.2010.536402

Balazünbül, Zeynep (2021): Drohnentechnologie und moderne Kriegführung. Bestandsaufnahme und Perspektiven. Berlin: Carl Grossmann. https://doi.org/10.24921/2021.94115956

Chesterman, Simon (2020): Artificial intelligence and the limits of legal personality. In: International and Comparative Law Quarterly 69 (4), pp. 819–844. https://doi.org/10.1017/S0020589320000366

Defence Ethics Committee (2021): Opinion on the integration of autonomy into lethal weapon systems. Paris: Ministère des Armées. Available online at https://cd-geneve.delegfrance.org/IMG/pdf/defence_ethics_committee_-_opinion_on_the_integration_of_autonomy_into_lethal_weapon_systems.pdf?2423/17d8f6beb2f5c9caa9c9168c53c24a91d9d32513, last accessed on 14. 05. 2024.

Department of Defense (2023): DoD Directive 3000.09. Autonomy in weapon systems, 25. 01. 2023. Available online at https://www.esd.whs.mil/portals/54/documents/dd/issuances/dodd/300009p.pdf, last accessed on 29. 04. 2024.

Dombrowsky, Wolf (2013): Risiko und Risikoprävention. In: Monatsschrift für Kriminologie und Strafrechtsreform 2/3, pp. 234–240. https://doi.org/10.1515/mks-2013-962-317

Dremliuga, Roman; Prisekina, Natalia (2020): The concept of culpability in criminal law and AI Systems. In: Journal of Politics and Law 13 (3), pp. 256–262. https://doi.org/10.5539/jpl.v13n3p256

Fabiano, Joao (2023): Should weaponised moral enhancement replace lethal aggression in war? In: Israel Law Review 56 (2), pp. 201–224. https://doi.org/10.1017/S002122372200022X

Gless, Sabine (2017): Von der Verantwortung einer E-Person. In: Goltdammer's Archiv für Strafrecht 164 (6), pp. 324–32. https://doi.org/10.5451/unibas-ep55627

Gless, Sabine; Weigend, Thomas (2015): Intelligente Agenten und das Strafrecht. In: Zeitschrift für die gesamte Strafrechtswissenschaft 126 (3), pp. 561–591. https://doi.org/10.1515/zstw-2014-0024

Guha, Manabrata; Galliott, Jai (2023): Autonomous systems and moral de-skilling. Beyond good and evil in the emergent battlespaces of the twenty-first century. In: Journal of Military Ethics 22 (1), pp. 51–71. https://doi.org/10.1080/15027570.2023.2232623

Haugstvedt, Håvard (2023): A flying reign of terror? The who, where, when, what, and how of non-state actors and armed drones. In: Journal of Human Security 19 (1), pp. 1–7. https://doi.org/10.12924/johs2023.19010001

ICRC – International Committee of the Red Cross (2014): Autonomous weapon systems, technical, military, legal and humanitarian aspects. Geneva: International Committee of the Red Cross. Available online at https://www.icrc.org/en/download/file/1707/4221-002-autonomous-weapons-systems-full-report.pdf, last accessed on 29. 04. 2024.

ICRC (2021): ICRC position on autonomous weapon systems. Geneva: International Committee of the Red Cross. Available online at https://www.icrc.org/en/download/file/166330/icrc_position_on_aws_and_background_paper.pdf, last accessed on 29. 04. 2024.

Lieblich, Eliav (2012): Show us the films. Transparency, national security and disclosure of information collected by advanced weapon systems under international law. In: Israel Law Review 45 (3), pp. 459–491. https://doi.org/10.1017/S0021223712000155

Meron, Theodor (2011): The making of international criminal justice. A view from the bench. Selected speeches. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199608935.001.0001

Mozur, Paul; Hopkins, Valerie (2023): Ukraine's war of drones runs into an obstacle – China. In: New York Times, 30. 09. 2023. Available online at https://www.nytimes.com/2023/09/30/technology/ukraine-russia-war-drones-china.html, last accessed on 29. 04. 2024.

Olásolo, Héctor (2005): The triggering procedure of the International Criminal Court. Boston, MA: Martinus Nijhoff. https://doi.org/10.1163/9789047415749

People's Republic of China (2022): Working paper on lethal autonomous weapons systems, 09. 08. 2022. Geneva: Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons System. Available online at https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-_Group_of_Governmental_Experts_(2022)/CCW-GGE.1-2022-WP.6.pdf, last accessed on 06. 05. 2024.

Porter, Jean (2010): Ministers of the law. A natural law theory of legal authority. Grand Rapids, MI: Wm. B. Eerdmans Publishing. https://doi.org/10.1111/j.1468-0025.2011.01733.x

Rabi, Shahaf; Plaw, Avery (2020): Israeli compliance with legal guidelines for targeted killing. In: Israel Law Review 53 (2), pp. 225–258. https://doi.org/10.1017/S0021223720000059

Reichert, Jochem (2011): Reaktionspflichten und Reaktionsmöglichkeiten der Organe auf (möglicherweise) strafrechtsrelevantes Verhalten innerhalb des Unternehmens. In: Zeitschrift für Internationale Strafrechtsdogmatik 6 (3), pp. 113–122. Available online at https://zis-online.com/dat/artikel/2011_3_535.pdf, last accessed on 29. 04. 2024.

Rutherford, Tim (2015): Everyone's accountable. How non-state armed groups interact with international humanitarian law. In: Australian Defence Force Journal 198, pp. 76–82. Available online at https://search.informit.org/doi/pdf/10.3316/ielapa.710945148772352, last accessed on 06. 05. 2024.

STO – Science & Technology Organization (2020): Science and technology trends 2020–2040. Brussels: NATO. Available online at https://www.sto.nato.int/publications/Management%20Reports/2020_TTR_Public_release_final.pdf, last accessed on 29. 04. 2024.

Simmler, Monika; Markwalder, Nora (2017): Roboter in der Verantwortung? In: Zeitschrift für die gesamte Strafrechtswissenschaft 129 (1), pp. 20–47. https://doi.org/10.1515/zstw-2017-0002

Steinberg, Gerald; Herzberg, Anne (2018): NGO fact-finding for IHL enforcement. In search of a new model. In: Israel Law Review 51 (2), pp. 261–299. https://doi.org/10.1017/S0021223718000079

Thompson, Kristen (2024): How the drone war in Ukraine is transforming conflict. In: Council of Foreign Relations, 16. 01. 2024. Available online at https://www.cfr.org/article/how-drone-war-ukraine-transforming-conflict, last accessed on 29. 04. 2024.

Trusilo, Daniel (2023): Autonomous AI systems in conflict. Emergent behavior and its impact on predictability and reliability. In: Journal of Military Ethics 22 (1), pp. 2–17. https://doi.org/10.1080/15027570.2023.2213985

UN – United Nations (1977): Protocol additional to the Geneva Conventions of 12 August 1949, and relating to the protection of victims of international armed conflicts (Protocol I) of 8 June 1977. Available online at https://www.un.org/en/genocideprevention/documents/atrocity-crimes/Doc.34_AP-I-EN.pdf, last accessed on 29. 04. 2024.

USAID – United States Agency for International Development (2023): USAID delivers Skydio autonomous camera drones to Ukraine to document war crimes. In: USAID Press Releases, 27. 07. 2023. Available online at https://www.usaid.gov/news-information/press-releases/jul-27-2023-usaid-delivers-skydio-autonomous-camera-drones-ukraine-document-war-crimes, last accessed on 29. 04. 2024.

Veilleux-Lepage, Yannick; Archambault, Emil (2022): A comparative study of non-state violent drone use in the Middle East. The Hague: ICCT. Available online at https://www.icct.nl/publication/comparative-study-non-state-violent-drone-use-middle-east, last accessed on 16. 05. 2024.

Wall, Tyler; Monahan, Torin (2011): Surveillance and violence from afar. The politics of drones and liminal security-scapes. In: Theoretical Criminology 15 (3), pp. 239–254. https://doi.org/10.1177/1362480610396650

DR. VASILIOS TZOUFIS
is a lawyer (Athens Bar Association) at the Supreme Court of Greece ("Areios Pagos") and a postdoctoral researcher at Aristotle University of Thessaloniki (Faculty of Law). He has taught at schools of the Greek army and the Greek security forces.

DR. NIKOLAOS PETROPOULOS
is a criminologist with a Ph. D. in criminal justice from the City University of New York. He is also a member of the EENeT Steering Committee.

Nachhaltigkeit
A–Z

O wie Openness

Welchen Wandel bedeuten Open Source, Open Government oder die Blockchain-Technologie? Diese Studie beleuchtet aus kultursemiotischer Sicht die Hintergründe des digitalen Sharings: von den Ursprüngen der Hackerbewegung über konkrete Anwendungen im Bankwesen oder Software bis hin zu demokratischem Diskurs und neuen Beteiligungsmöglichkeiten.

P. Dobroć
Open Source, Open Government, Blockchain
Eine kultursemiotische Studie zur Openness-Vision
294 Seiten, Broschur, 36 Euro
ISBN 978-3-98726-056-8

Bestellbar im Buchhandel und unter www.oekom.de.
Auch als E-Book erhältlich.

Die guten Seiten der Zukunft
oekom

TATuP Zeitschrift für Technikfolgenabschätzung in Theorie und Praxis
Journal for Technology Assessment in Theory and Practice

oekom science

RESEARCH ARTICLE

Perception of the risks inherent in new AI technologies

Petr Machleidt*,[1], Jitka Mráčková[2], Karel Mráček[3]

Abstract • Artificial intelligence (AI) has undergone rapid development and is becoming one of the major social issues. The advent of generative AI is not only associated with potential benefits but also poses a number of risks, such as increasing malicious misuse. This brings the issue of regulating new AI technologies into focus. In the search for solutions, the article revisits the problem of applying the precautionary principle. We critically evaluate regulatory approaches, particularly with regard to maintaining an innovation-friendly environment. A prudent approach to new AI technologies not only requires regulatory measures but also places new demands on the education system. We also discuss the regulation of AI from the perspective of Czech legislation in a broader international context. Current challenges in the field of AI competence are highlighted. To make the issue more tangible, we provide specific examples from the Czech Republic.

Wahrnehmung der mit neuen KI-Technologien verbundenen Risiken

Zusammenfassung • *Künstliche Intelligenz (KI) entwickelt sich rasant weiter und wird mehr und mehr zu einem zentralen gesellschaftlichen Thema. Das Aufkommen der generativen KI ist nicht nur mit potenziellen Vorteilen verbunden, sondern birgt auch Risiken, wie etwa einen zunehmenden böswilligen Missbrauch. Damit rückt die Frage der Regulierung neuer KI-Technologien in den Fokus. Auf der Suche nach Lösungsansätzen für diese Herausforderungen geht der Artikel erneut auf die Problematik der Anwendung des Vorsorgeprinzips ein. Regulierungsansätze werden vor allem mit Blick auf die Aufrechterhaltung eines innovationsfreundlichen Umfelds kritisch bewertet. Ein umsichtiger Umgang mit neuen KI-Technologien erfordert nicht nur regulatorische Maßnahmen, sondern stellt auch neue Anforderungen an das Bildungssystem. Darüber hinaus wird im Artikel die Regulierung von KI aus Sicht der tschechischen Gesetzgebung im breiteren internationalen Rahmen diskutiert. Dabei werden auch aktuelle Herausforderungen im Bereich der KI-Kompetenz beleuchtet. Um die Problematik greifbarer zu machen, werden konkrete Beispiele aus der Tschechischen Republik angeführt.*

Keywords • *artificial intelligence, regulation, ethics, risks, AI Act, education*

This article is part of the Special topic "Malevolent creativity and civil security: The ambivalence of emergent technologies," *edited by A. Gazos, O. Madeira, G. Plattner, T. Röller, and C. Büscher. https://doi.org/10.14512/tatup.33.2.08*

Introduction

The new challenge for technology assessment is the rapid development of artificial intelligence (AI), which is unprecedented in its history to date. AI technologies penetrate various fields of human activity, open tremendous opportunities, but also risks in the form of harmful use or misuse. Consequently, AI involves a variety of security risks, further compounded by its current rapid development.

Malevolent use of new AI technologies is a problem perceived in the Czech Republic as well. How to solve this new technological and social challenge? What can we do about it? The application of the precautionary principle is offered as a framework for regulatory measures to achieve trustworthy AI. However, a precautionary approach must also be supported by educational efforts in AI literacy. These approaches should be critically evaluated with a view to preserve a pro-innovation environment.

Views on AI risks

In debates across the media in the Czech Republic, the biggest civil-security risks of AI associated with its use for malevolent

* *Corresponding author: machleidt@flu.cas.cz*
[1] *Institute of Philosophy, Czech Academy of Sciences, Prague, CZ*
[2] *Czech University of Life Sciences Prague, Prague, CZ*
[3] *Czech Technical University in Prague, Prague, CZ*

© 2024 by the authors; licensee oekom. This Open Access article is licensed under a Creative Commons Attribution 4.0 International License (CC BY).
https://doi.org/10.14512/tatup.33.2.42
Received: 04. 01. 2024; revised version accepted: 24. 04. 2024;
published online: 28. 06. 2024 (peer review)

OPEN ACCESS

activities are perceived in the area of personality protection and its privacy and the space created for the spread of disinformation, consumer protection and in the healthcare sector (sensitive data).

When updating the national strategy for artificial intelligence, a public consultation was conducted to source stakeholders' and the wider public's views of the opportunities and threats presented by the contemporary dynamic advancement of AI (MI-ToCR 2023). As regards the perceived threats, most respondents (about 62 %) mentioned manipulation and proliferation of fake information. Other AI-related threats, by frequency of occurrence, included cyber-attacks, disappearance of human agency in the decision-making process, technology dependence, invasion of privacy, digital divide, lack of regulation and loss of jobs. In general, in the respondents' opinion, the most important issues include education, the use of AI in practice and the creation of a legal framework (which would lay down a clear set of rules and ensure the protection of citizens' rights).

Certain concerns were also voiced in earlier research focusing on AI in financial services with implications for the Czech population (Ipsos 2023). Three quarters of all the respondents believe that the use of AI in financial services should be curbed by regulation (in the form of principles and rules). In terms of privacy and security of financial data that involve the use of AI, more than one third of the respondents feel safe while one quarter of them do not. However, it is precisely financial services that give rise to the interesting phenomenon of AI technology ambivalence where the threat of malicious use of AI (in the form of phishing and other cyber fraud) is offset by AI's potential for rapid detection and even elimination of such incidents. One statistic to be mentioned in this respect is the rate of confidence in AI's role in fraud detection and prevention (36 % of respondents).

The advent of new generative AI (GenAI) technologies in the Czech Republic in the months of 2023 that followed manifested itself in particular in a surge in digital attacks and fraud attempts targeting clients of banks and other financial sector institutions as a chronically weak link in financial security. The use of new GenAI technologies makes it possible to create fraudulent emails, fake websites and deepfake videos, fake photos and voices faster, in greater quantities and at higher quality levels than has been possible so far. While, previously, Czech clients were to some extent protected by the complexity of the Czech language, this is no longer the case with the arrival of AI's new generation. In finance, the main concern is the misuse of voice in soliciting non-standard operations. Financial institutions frequently use chatbots to communicate with their clients, so the latter may often fail to recognise that they are being scammed. In the Czech Republic, we have recently seen a surge in a number of deep fake videos – a traditional tool in creating disinformation campaigns with a view to swaying public opinion for political gains – used for financial fraud purposes. The videos usually feature a celebrity promoting fake investment opportunities.

According to statistics, cybercrime in the Czech Republic is steadily increasing and in 2023 it accounted for approx. 11 % of the total recorded crime (Moravčík and Vinčálek 2024). It is especially more sophisticated phishing practices and recruitment of money launderers on social media and other online platforms that have become more prevalent following the emergence of GenAI. According to the National Cyber and Information Security Agency (NÚKIB), the number of cyber incidents involving critical infrastructure doubled year-on-year in 2023, with security experts linking the trend mainly to the arrival of more sophisticated AI technologies (NÚKIB 2024).

While, previously, Czech clients were to some extent protected by the complexity of the Czech language, this is no longer the case with the arrival of AI's new generation.

According to the AI Trends survey (Randstad HR Solutions s.r.o. 2023), two thirds of companies assume AI requires a legal framework, 28 % of them believe that the regulation should be very strict. The corporate sector is generally more reticent to use GenAI compared to the general population. Some of the concerns voiced include opening corporate data to a global AI service for the latter to learn from, and the resulting threat of corporate know-how being simply handed over to just about anybody. At the same time, there is a growing awareness of the potential of AI in data protection and a range of its applications in cyber protection. Overall, the situation can be described as moderate techno-optimism coupled with a certain amount of caution.

The analyses and investigations conducted show a consensus on the need for adequate legal regulation and education in the development and use of AI.

Background of AI regulation

The rapidly evolving AI and the associated and increasingly perceived risk of its malevolent use call for attention to the issue of regulating novel technologies. As a result, the way the precautionary principle – no matter how ambivalently accepted – is applied needs to be revisited. The precautionary principle has its proponents, but it is also subject to criticism. The proponents generally view the principle as an important concept to protecting the environment and human health (Randall 2011). In partic-

https://doi.org/10.14512/tatup.33.2.42

ular, they argue that the application of the principle contributes to the development of safe technologies and operates as means of protection against the potential risks of new technology and products. As such, the precautionary principle is closely related to risk management and is usually applied in situations where the potential risks have been demonstrated, yet some uncertainty remains as to their extent or severity.

The critics regard the principle as an obstacle to further innovation, stressing its potential for over-regulation (Sunstein 2002). Excessive caution can indeed lead to disproportionate requirements for evidence demonstrating the safety of new technologies and products, thus stifling the innovation process. The focus then shifts away from the potential benefits of new technologies to the risks inherent in them. When applied, the precautionary principle can lead to indecision due to a lack of sufficient evidence, which can have a negative impact on the economy and society at large. Another aspect criticised in this context is to overload a person with excessive responsibility, which can have a demotivating effect on him. In terms of international trade, there are concerns involving the misuse of the principle for protectionism (Marchant and Mossman 2004).

In the Czech Republic, the precautionary principle is included among the principles governing the environmental protection policy and is defined in the Environmental Act.

The requirement to preserve a pro-innovation environment is prominently featured in current discussions on AI regulation (OECD 2023). Regulation that is too restrictive and allows little flexibility may stifle innovation. Arguably, it is in the highly dynamic field of AI that the objectives of regulation should be primarily set in the form of fundamental rights, protection of human health and safety and the environment, in conjunction with the acceptable rules for AI's trustworthiness (such as transparency, protection of the personality and of privacy, and combating discrimination); and individuals or firms should be left to come up with their own technical solutions to meet these requirements. Technical standards that specify a certain level of attainment of an objective and various test facilities are also useful in this context. Statutory regulation may stifle innovation if it is overly convoluted and ambiguous, allowing for different interpretations. Regulation that leads to a disproportionate growth in initial costs and an increase in administration may hinder innovation just as well. This situation can be particularly burdensome for SMEs and start-ups. The dynamic development of modern technologies – and this especially applies to AI given its extreme rate of advancement – presents the law with the problematic task of constantly having to catch up with the developments and reduces the legislator's role to merely adapting the legal framework to those developments. The first version of the AI Act was released in 2021, and it was only after that generative AI models emerged, creating new requirements for legislators to adapt the upcoming regulation accordingly. This also raises the provocative question as to whether we want to regulate something we know little with regard to the non-transparency of the inner workings of the newly developing AI models.

Yet, what clearly legitimises regulation of innovations is the need to prevent potential risks inherent in the use or even misuse of such novelties. Where innovations may create negative externalities, appropriate regulation of such externalities is entirely justified. In this context, the precautionary principle indeed seems to be an appropriate instrument of containing potential hazards that may not even be known (information asymmetry). The instances where the general precautionary principle has so far been applied are related mainly to the environmental segment and to food safety and pharmaceuticals. Germany played a major role in laying the philosophical and ethical foundations of the precautionary principle and in the way it was eventually embedded in policy and law. The European understanding of the principle in EU sources of law is inspired by German environmental law (Kühn 2006, p. 492). In the Czech Republic, the precautionary principle is included among the principles governing the environmental protection policy and is defined in the Environmental Act.

In Czech legal literature, the issue is comprehensively dealt with only from the viewpoint of international commercial law (Grmelová 2022). The EU regards the precautionary principle as a general legal principle applicable to all areas of law.[1] While the AI Act does not explicitly refer to this principle, it does apply a certain precautionary tenet to AI in the form of a risk-based approach. In essence, this reflects the universal recognition of the need to protect valuable goods such as the environment, life and human integrity amidst concerns induced by unregulated technological development on a global scale. GenAI is concerned with co-creation and, by inference, creativity, and it is expected to interact with humans in working and creative environments, but, as the obtained knowledge suggests, it can also be exploited for malevolent creativity. The impacts of positive use or misuse are much more far-reaching and global for this technology, and are associated with a range of uncertainties concerning future developments. Consequently, the requirement for precautionary measures, particularly as regards the safeguarding of moral

1 EC, Artegodan v. Commission, Judgment of 26. November 2002, Case T-74/00 DEP.

https://doi.org/10.14512/tatup.33.2.42

rights and protection of privacy and intellectual property, is entirely justified. However, the principle should only be applied where necessary and in a proportionate manner where a lack of information and the potential risk persist after an evaluation based on the available scientific knowledge and the technology assessment carried out.

Regulatory measures and actions for trustworthy AI

Potentially different regulatory frameworks are currently applied in the world to ensure trustworthy AI and mitigate the risks associated with the development and deployment of AI systems. In addition to exploring ways to use and adapt current legislation for AI, new regulatory measures and actions to ensure trustworthy AI are coming to the fore in the form of (i) ethical frameworks and principles, (ii) hard law approaches, (iii) promoting international standardisation and international law efforts, and (iv) supporting the creation of controlled environments for regulatory experimentation (OECD 2023).

Some countries (Australia, Singapore, Switzerland, South Korea and others) have released national ethical frameworks for the development and deployment of AI in line with the OECD AI Principles (OECD 2023). The European Commission released its own ethics guidelines for the trustworthy use of AI in April 2019 (EC 2019). As AI penetrates still deeper into everyday use, the discussion on the ethics of AI, respecting the possibilities and limits of the approach, is becoming increasingly important. This, above all, concerns the ethical approaches adopted by firms developing and implementing AI. While ethical principles are not legally enforceable and operate on a voluntary basis, they contribute to the cultivation of the corporate and societal environment.

As regards AI regulation based on legislation, a distinction must be made between efforts at taking a cross-sectoral, 'horizontal' approach to such regulation (e.g. in the EU, Canada), and a context-based sectoral or 'vertical' approach where legislation focuses on sectors or areas (e.g., the US, UK, China, Israel). It should be borne in mind that specific legislation differs when it comes to the actual definition of AI, risk classification, and the scope and oversight of regulation. The definition of AI itself is a frequent subject of debates considering its importance in terms of the future scope and extent of the regulation.

In contrast to the EU, the legislators in the US are currently far from even considering an approach to AI regulation that would entail comprehensive legislation. Binding federal technology legislation is mostly sector or domain specific. Executive policy is important in that context (The White House 2022; NIST 2023). The Office of the President also drew voluntary commitments from leading U.S. technology companies to manage the risks associated with AI and to assist in the transition to safe, transparent, and trustworthy development of AI technologies. At the same time, AI risk management is linked to the promotion of responsible innovation (The White House 2023).

The UK's approach to AI regulation also differs significantly from the EU AI Act. It is characterised by devolution and focus on actual risks and harms (and on risk illustration rather than classification); it relies on sectoral regulation and favours voluntary measures and guidance. The regulatory option is a vertical approach with an emphasis on a pro-innovation framework (DSIT 2023). Regulators are advised to favour soft law over mandatory regulation and are left to set out rules specific to each sector.

The AI Act is based on a risk-assessment approach and provides a single horizontal legal framework for AI with a view to providing legal certainty and, avoiding fragmented regulation across EU Member States (EC 2023). The Act uses a fairly broad definition of AI, rating different AI systems on the scale of the risks they pose to society and risks are classified as unacceptable, high, limited and low or minimal. AI systems with an unacceptable level of risk to human safety will be prohibited (see, for instance, use of 'social scoring'). However, more accurate definitions of the different risk systems are also a challenge, taking into account the difficulty of enforcement, certification and addressing accountability issues. In addition, as AI permeates various fields of human activity, it is becoming a challenge for a number of legal sectors.

The dynamic developments in AI present another challenge for the rather complex EU legislative process ('the trialogue'). The original text of the AI Act had to respond to major changes (see, for instance, GenAI). The looming dilemma of the rampant development of AI and delays in the legislative process are somewhat reminiscent, in retrospect, of the inadequate regulation problem concerning social networks algorithms (see the so-called Collingridge dilemma).

In the Czech Republic, much attention has been paid in recent years to the ethical and legal aspects of AI (AVex 2023). The Czech legal system is influenced by the developments in AI legislation in the EU and international law. In designing downstream AI regulation, more inspiration can be drawn from the precautionary principle in 'Leitlinien zur Umweltvorsorge' from 1986, where three sub-objectives are identified (Arndt 2009). The objectives may be put to good use in the process of regulating AI. Going forward, the advancement and use of AI will bring about the need to avert harm to humans, mitigate risk and act prudently. The precautionary principle is contained in the aforementioned Czech Act No. 17/1992 Coll., on the environment (Section 13), according to which the principle may be applied *per analogiam* in order to close a gap in the law. AI raises a number of outstanding questions that will have to be addressed across the entire Czech law. As for the adopted AI Act, the Czech Republic is now in the process of familiarising itself with the legislation. The oversight authority issue will need to be resolved and the adopted European rules will have to be transposed into national legislation (in particular civil, criminal and administrative law).

In civil law, AI is viewed as a sui generis intangible thing, i.e. an object (as opposed to a subject) of a legal relationship, which

https://doi.org/10.14512/tatup.33.2.42

removes any doubt as to the possibility of granting legal personality to AI (Zibner 2022). According to the Czech Copyright Act, which is based on the principle of objective authorship, only a natural person can be an author as the originator of a work. AI cannot be an author even in common law countries because it is not a human being. In civil law systems, an author (a human being) retains their moral and economic rights under copyright during their lifetime. Only the right to use the copyrighted work can be transferred (the licensee being a mere user, as opposed to owner, of the work). The same applies to AI, and this may raise some concerns when addressing liability for material and non-material damage. Pending the development of appropriate AI legislation, the relevant cases will have to be resolved by way of analogy. Attention needs to be given in particular to the responsible entities in the development and use/misuse of AI (developers, programmers or, where applicable, employers, distributors, operators or users) with regard to the damage caused to individuals and legal persons, or the State. The rights and duties of these entities remain the subject of some discussion instead of being subjected to an in-depth analysis of the existing legal concepts of liability. Knapp, for example, tackled the issue of civil liability asking whether the subject is responsible for fulfilling a duty or only liable for its non-fulfilment, or a breach a duty. The answer to the question is that "whoever owes a duty is liable at once for the fulfilment of the duty and for a breach thereof". He went on to say that "they are responsible for the fulfilment of a duty in the sense that from its very onset, the law threatens them with a sanction in case they fail to fulfil the duty, and they are liable for its non-fulfilment in the sense that they will suffer a sanction as consequence of their liability" (Knapp 1956, both quotations p. 75). This approach connects a continuously prospective and retrospective concept of responsibility and liability, which could lead to the fulfilment of the precautionary principle. A similar approach can be found in German legal theory (Medicus and Lorenz 2008, p. 9). These theoretical concepts could be of use in regulating AI along its entire value chain (from development to use). It would be appropriate to establish risk mitigation obligations for the different actors in the chain and make them liable for fulfilling the obligations, not only for breaching them. And it is precisely such coupling of the prospective and retrospective liability that could be found operative in relation to AI-relevant legal acts. This will require flexible monitoring mechanisms. Statutory regulation must be effective and allow for adequate oversight in order to enhance legal certainty.

Education activities related to AI

A precautionary approach in dealing with new technologies does not necessarily have to rely on regulatory measures; the education system must also play a role in this. The latter is now gaining extra importance as the rapid development of GenAI leads to significant changes in the labour market. While in the past novel technologies have traditionally had the effect of eliminating routine and physically demanding work, the present changes will also affect a wide range of creative professions. This calls for increased competences of the general public to work with AI. AI literacy therefore poses a tremendous challenge to the entire education system, ranging from school education to lifelong learning. AI literacy must include the skills to harness the opportunities and minimise the negative impacts of AI, to understand the possibilities and limits of these technologies. In terms of precautionary measures, particular attention must be devoted to the ways of identifying and interpreting dangerous uses of AI in the form of disinformation, manipulation and deepfakes and of defending against such manipulative techniques. Malevolent AI users are not only using technical innovations in this area; recently they have innovated the way of social communication and manipulation of victims. The risk of the pool of competent persons capable of malevolent AI creativity deepening must also be taken into account. AI literacy can work more effectively as means of protection against malevolent AI use, if combined with media literacy, financial literacy, etc., so that synergies in the competences achieved can be exploited. AI literacy also includes an understanding of the ethical aspects of the development and use of AI. Organisations should foster a culture of ethical use of AI, educating employees on the ethical aspects of AI (such as bias, transparency and privacy).

It would be appropriate to establish risk mitigation obligations for the different actors in the chain and make them liable for fulfilling the obligations, not only for breaching them.

So far, the Czech Republic has predominantly relied on the bottom-up approach to AI skill acquisition. Noteworthy is the upsurge in various educational, awareness-raising and support activities. In 2023, for instance, the AI Weeks event was held; PRG.AI and Brno.AI, the two associations established with a view to promoting regional research, education, innovation and entrepreneurship, have become fully operational; and media attention to AI has increased manifold. Businesses and various organisations are offering AI-focused educational events for their staff. Due to the ever-increasing threat of deepfake attacks on businesses, Analytics Data Factory, in cooperation with the Czech Association of Artificial Intelligence, has created a free manual entitled 'DEEPFAKE 2024': Defence Strategy for

 https://doi.org/10.14512/tatup.33.2.42

Czech Companies (CAUI 2024). To train their staff, the largest banks on the Czech market use internal versions of ChatGPT, which provide access only to public information.

The rapid spread of ChatGPT brought a wave of interest in AI at all levels of the education system. Related initiatives have already been put in place at many schools. However, further developments will not be possible unless ground rules are set, guidance is provided and certain ethical issues are resolved (e.g., ensuring the privacy and security of students' data). The use of GenAI in higher education raises challenging questions and views. GenAI is highlighted as a source of inspiration (offering new topics, different perspectives, out-of-the-box solutions). It is also well suited for generating various tutorials, manuals, informative texts and statistics, but these may not necessarily reflect the views and attitudes of the author and their personality (Kopecký 2023).

Conclusion

New AI technologies bring a score of opportunities and benefits, but also risks, including the danger of increasing malevolent misuse. Possible approaches to minimizing this perceived threat are mainly in adequate regulatory measures and educational activities. In terms of the regulations, AI development and use have mostly been regulated by ethical rules. It now seems desirable to evaluate these rules in terms of their efficacy and to consider which of them should be given greater force in the form of statutory law. From the viewpoint of AI legislation, we need to address important questions concerning the various roles of law (regulatory and supervisory, and legal certainty roles). There are certain yet unresolved questions as to the rights and duties relating to AI as an intangible thing that pertain to subjects of legal relations (software originators, users, etc.).

A prudent approach to new technologies does not have to consist only of regulatory measures, but the education system must also play its role. We must always keep in mind that "technology development should not be left to random coincidences and to people who use the technology regardless of its side effects and who do not consider the prudence aspect in the application of completely new means" (Drozenová 2019, p. 65).

A coordinated approach and cooperation between policy makers, state bodies, researchers and businesses (especially big tech companies) is proving indispensable. Investments in further research and greater support for lifelong learning in this area clearly have a role to play in increasing resilience to malevolent AI use. Overall, a proactive approach is essential if we want to maximise the potential of AI while minimising its risks. Considering the need for the formation of socially acceptable approaches to modern disruptive technologies, keeping AI under human control is a major challenge.

Funding • This work received no external funding.

Competing interests • The authors declare no competing interests.

References

Arndt, Birger (2009): Das Vorsorgeprinzip im EU-Recht. Tübingen: Mohr Siebeck.

AVex – Akademie věd České republiky (2023): Umělá inteligence. Available online at https://www.avcr.cz/export/sites/avcr.cz/cs/veda-a-vyzkum/avex/files/2023-01.pdf, last accessed on 15. 05. 2024.

CAUI – Česká asociace umělé intelligence (2024): Deepfake 2024. Obranná strategie pro české firmy. Available online at https://asociace.ai/wp-content/uploads/2023/12/DEEPFAKE-2024-CAUI.pdf, last accessed on 26. 04. 2024.

Drozenová, Wendy (2019): Jonasova etika techniky a její metafyzické ukotvení. In: Wendy Drozenová and Vojtěch Šimek (eds.): Filosofie Hanse Jonase. Praha: Filosofia, pp. 61–79.

DSIT – Department for Science, Innovation and Technology (2023): A pro-innovation approach to AI regulation. Available online at https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach, last accessed on 26. 04. 2024.

EC – European Commission (2019): Ethics guidelines for trustworthy AI. Available online at https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai, last accessed on 26. 04. 2024.

EC (2023): European approach to artificial intelligence. Available online at https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence, last accessed on 26. 04. 2024.

Grmelová, Nicole (2022): Zásada předběžné opatrnosti v právu mezinárodního obchodu. Praha: C. H. Beck.

Ipsos (2023): AI ve finančních službách. Available online at https://www.ipsos.com/cs-cz, last accessed on 15. 05. 2024.

Knapp, Viktor (1956): Některé úvahy o odpovědnosti v občanském právu. In: Stát a právo 1, pp. 66–85.

Kopecký, Kamil (2023): Využití AI ve vysokoškolském vzdělávání. In: Kamil Kopecký Blog, 03. 09. 2023. Available online at https://kopeckykamil.cz/index.php/blog/351-vyuziti-umele-inteligence-ve-vysokoskolskem-vzdelavani-podpora-studentu, last accessed on 26. 04. 2024.

Kühn, Werner (2006): Die Entwicklung des Vorsorgeprinzips im Europarecht. In: Zeitschrift für europarechtliche Studien 9 (4), pp. 487–520. https://doi.org/10.5771/1435-439X-2006-4-487

Marchant, Gary; Mossman, Kenneth (2004): Arbitrary and capricious. The Precautionary Principle in the European Courts. Washington, DC: American Enterprise Institute for Public Policy Research. Available online at https://www.aei.org/wp-content/uploads/2011/11/20040917_MarchantNewG.pdf?x85095, last accessed on 16. 05. 2024.

Medicus, Dieter; Lorenz, Stephan (2008): Schuldrecht I. Allgemeiner Teil. München: C. H. Beck.

MIToCR – Ministry of Industry and Trade of the Czech Republic (2023): Vyhodnocení veřejné konzultace. Available online at https://www.mpo.cz/assets/cz/podnikani/digitalni-ekonomika/umela-inteligence/2023/10/Vyhodnoceni-verejne-konzultace-k-aktualizaci-Narodni-strategie-umele-inteligence.pdf, last accessed on, 26. 04. 2024.

Moravčík, Ondřej; Vinčálek, Jakub (2024): Vývoj registrované kriminality v roce 2023. In: Policie České republiky, 12. 01. 2024. Available online at https://www.policie.cz/clanek/vyvoj-registrovane-kriminality-v-roce-2023.aspx, last accessed on 24. 04. 2024.

NIST – National Institute of Standards and Technology (2023): Artificial intelligence risk management framework (AI RMF 1.0). Washington, DC: U.S. Department of Commerce. https://doi.org/10.6028/NIST.AI.100-1

https://doi.org/10.14512/tatup.33.2.42

NÚKIB (2024): NÚKIB v roce 2023 zaznamenal rekordní počet kybernetických incidentů, 31. 01. 2024. Available online at https://nukib.gov.cz/cs/infoservis/aktuality/2073-nukib-v-roce-2023-zaznamenal-rekordni-pocet-kybernetickych-incidentu/, last accessed on 26. 04. 2024.

OECD – Organization for Economic Co-operation and Development (2023): The state of implementation of the OECD AI principles four years on. OECD Artificial Intelligence Papers 3. Paris: OECD Publishing. https://doi.org/10.1787/835641c9-en

Randall, Alan (2011): Risk and precaution. Cambridge, UK: Cambridge University Press.

Randstad HR Solutions s.r.o. (2023): Průzkum AI Trends 2023: Firmy kvůli umělé inteligenci propouštět neplánují. In: Randstad.cz, 26. 10. 2023. Available online at https://www.randstad.cz/o-nas/randstad-employer-brand-research/pruzkum-ai-trends-2023-firmy-kvuli-umele-inteligenci/, last accessed on 26. 04. 2024.

Sunstein, Cass (2002): Risk and reason. Safety, law, and the environment. Cambridge, UK: Cambridge University Press.

The White House (2022): Blueprint for an AI Bill of Rights. Available online at https://www.whitehouse.gov/ostp/ai-bill-of-rights, last accessed on 16. 05. 2024.

The White House (2023): President Biden issues Executive Order on safe, secure, and trustworthy artificial intelligence, 30. 10. 2023. Available online at https://www.whitehouse.gov/briefing-room/statements-releases/2023/10/30/fact-sheet-president-biden-issues-executive-order-on-safe-secure-and-trustworthy-artificial-intelligence/, last accessed on 26. 04. 2024.

Zibner, Jan (2022): Umělá inteligence jako technologická výzva autorskému právu. Praha: Wolters Kluwer.

DR. PETR MACHLEIDT
is an associate researcher at the Centre for Science, Technology, and Society Studies of the Institute of Philosophy of the Czech Academy of Sciences. His research interest is technology assessment; member of the CULTMEDIA project; 1995–2014 he lectured at the Czech Technical University in Prague on the subject of social assessment of technology.

DR. JITKA MRÁČKOVÁ
is an assistant professor at the Law Department of the Czech University of Life Sciences in Prague; lecturing and research activities in civil law and commercial law, legal aspects of informatics with a focus on the protection of intellectual property, previously intellectual property law also at Academy of Performing Arts in Prague.

DR. KAREL MRÁČEK
sets his focus of research on technology and innovation assessment, innovation policy and management; participation in the CULTMEDIA project. He lectures since 2015 at the Czech Technical University in Prague (guarantees the subject of social assessment of technology). Member of the Board of the Association of Research Organizations of the Czech Republic.

Nachhaltigkeit
A–Z

P wie Pionierarbeit

Nachhaltigkeit und Digitalisierung sind die großen Megatrends unserer Zeit. Dieses Buch ist ein Wegweiser zu dieser doppelten Transformation. Es zeigt anhand von inspirierenden Good-Practice-Beispielen, wie Unternehmen Pionierarbeit für eine grünere Zukunft leisten.

J. Quaing, J. Fink, B. Bilfinger, F. Vorländer
Doppelte Transformation gestalten
Ein Praxisleitfaden zu Nachhaltigkeit und Digitalisierung
220 Seiten, Broschur, vierfarbig mit Infografiken, 36 Euro
ISBN 978-3-96238-129-5

Bestellbar im Buchhandel und unter www.oekom.de.
Auch als E-Book erhältlich.

Die guten Seiten der Zukunft
oekom

https://doi.org/10.14512/tatup.33.2.42

TATuP Zeitschrift für Technikfolgenabschätzung in Theorie und Praxis
Journal for Technology Assessment in Theory and Practice

oekom science

RESEARCH ARTICLE

Use case generation, variation, and anticipation in innovation projects: Analysis of current industry practices based on interviews with international innovation experts

Niklas Hermann Henke*, 1, 2

Abstract • Use case scenarios play an essential role in innovation projects and influence the design of technologies. In this research article, I examine how use cases are generated, varied, and anticipated. To do so, I analyze interviews with international innovation experts (N = 14), e.g., Chief Technology Officer, Innovation Director, or User Experience Researcher. A heterogeneity of practices for the design of use cases, the creation of use case variants, and the anticipation of a possible extension of use cases can be identified. Anti-personas and worst-case scenarios are tools specifically aimed at identifying malicious use cases. A surprising result is the absence of special AI-based procedures for use case generation, variation, and anticipation, indicating untapped potential for innovation projects.

***Generierung, Variation und Antizipation von Anwendungsfällen bei Innovationsprojekten**: Analyse aktueller Industriepraktiken anhand von Interviews mit internationalen Innovationsexperten*

Zusammenfassung • *Anwendungsszenarien spielen in Innovationsprojekten eine wesentliche Rolle und haben Einfluss auf die Gestaltung von Technologien. In diesem Forschungsartikel untersuche ich, wie Anwendungsfälle generiert, variiert und antizipiert werden. Dazu analysiere ich Interviews mit internationalen Innovationsexperten (N = 14), z. B. Chief Technology Officer, Innovation Director oder User Experience Researcher. Für die Gestaltung von Anwendungsfällen, die Erstellung von Anwendungsfallvarianten und die Antizipation einer möglichen Erweiterung von Anwendungsfällen lässt sich eine Heterogenität von Praktiken feststellen. Anti-Personas und Worst-Case-Szenarien sind Instrumente, die speziell auf die Identifizierung böswilliger Anwendungsfälle abzielen. Ein überraschendes Ergebnis ist das Fehlen spezieller KI-basierter Verfahren zur Generierung, Variation und Antizipation von Anwendungsfällen, was auf ein ungenutztes Potenzial für Innovationsprojekte hindeutet.*

Keywords • *innovation, creativity, use case, artificial intelligence, technology*

This article is part of the Special topic "Malevolent creativity and civil security: The ambivalence of emergent technologies," *edited by A. Gazos, O. Madeira, G. Plattner, T. Röller, and C. Büscher. https://doi.org/10.14512/tatup.33.2.08*

Introduction

Societal implications of innovations, especially technological ones, get ever more visible and cannot be underestimated. Innovation can be understood as "the implementation of a new or significantly improved product (good or service), or process, a new marketing method, or a new organizational method in business practices, workplace organization or external relations." (OECD and Eurostat 2005) The procedures, methods, and tools applied during innovation projects impact the final innovations, and con-

* *Corresponding author: niklas.henke@capgemini.com*
[1] *Capgemini Engineering, Toulouse, FR*
[2] *GRESEC laboratory, Université Grenoble Alpes, Grenoble, FR*

© 2024 by the authors; licensee oekom. This Open Access article is licensed under a Creative Commons Attribution 4.0 International License (CC BY).
https://doi.org/10.14512/tatup.33.2.49
Received: 03. 01. 2024; revised version accepted: 08. 04. 2024;
published online: 28. 06. 2024 (peer review)

OPEN ACCESS

sequently the daily lives of users (Falahat et al. 2024). Therefore, the industrial processes leading to innovations (so called 'innovation projects') gain growing societal attention and scientific interest. This leads to the specific role of use case generation as one of the building blocks of innovation concepts. This article looks at innovation projects from a communicational perspective, focusing on current communication and collaboration practices (Albastaki et al. 2022; Cicek 2024).

Use case generation means to describe a precise situation in which an innovation gets used, by whom, under which circumstances, to what benefit, and the emotions that motivate it, eventually including other information necessary for understanding the given innovation concept. Use cases are used to communicate an innovation concept to stakeholders and to test it with potential users. The involved teams aim at defining the innovation concept as precisely as possible and elaborating the probability of the innovation's larger diffusion (Rogers 1983).

There are certain usage extensions that are not only creative but malevolent. Socioeconomic actors show an interest in anticipating such cases as early as possible. The goal of this article is to identify cross-sectorial procedures to develop, vary, and anticipate use cases, and to understand their participation in the creation of new technologies. Industrial innovation practices undergo significant transformations, especially linked to AI tools. I expect to find manifestations of these trends via the chosen research design.

Research questions

How do innovation teams generate, vary, and anticipate use cases? Furthermore, which coping strategies are mobilized to address and avoid certain use cases (especially malevolent ones)? The first question is exploratory, aiming to identify any existing procedures, practices, methods, or tools, as well as personal preferences and opinions. The second question represents the logical consequence of the first. Among the variety of different use cases, some are classified as favorable and others as non-favorable by the innovation team. The question is intended to identify how innovation processes are set up to reinforce certain usages and avoid others. I aim to approach these questions in an exploratory manner to gain first insights into the transformation of innovation practices, and especially use case generation, via AI-based tools. Quantification and representativity shall be addressed in upcoming research.

Methodology

During October and November 2023, I conducted semi-structured interviews with highly-qualified, cross-sectorial, international innovation experts ($N = 14$). The profiles included Chief Technology Officer, UX Researcher, or Innovation Consultant in the automotive, consulting, chemical, IT, fashion, heavy manufacturing, or public services sectors. Focusing on a particular type of innovation project would have required fewer interviews, but I preferred a cross-sectorial approach to search for procedural patterns in a larger sense (Guest et al. 2006). The experts are based in Europe, Africa, North America, and the Middle East. The interviews were held in English, French, and German, took approximately one hour, and occurred online via Microsoft Teams. The experts were recruited through LinkedIn Sales Navigator and were not compensated for the interview. The interview guide aimed to understand the personal experiences and opinions of the experts instead of official procedures. It included four major sections: 1. Home Office, 2. Human Body, 3. Creativity, 4. Use Cases. The first three sections aimed at understanding the experts' working conditions, applied methods, creativity tools, and physical routines. The fourth section was focused on use case generation, variation, and anticipation. The first section included the perceived transition towards home offices, their ergonomic setups, collaboration tools, and relationships towards colleagues. The second section included physical habits, physical imaginaries/stereotypes of the user, and ergonomic norms/reference systems. The third section included the overall project setup, applied methods, and more specifically, creative methods/rituals. The fourth section included use case generation practices, procedures for identifying malevolent use cases, and the integration of AI-based tools. The interview was introduced via a short research description and technical information from the author. The experts were then asked to describe themselves, their professional roles, and backgrounds. The interview was concluded by asking for any last remarks, associations, ideas, or topics.

Results

The new normal – innovation via distance

The experiences of the innovation experts vary greatly. I can identify a heterogeneity linked to their personal circumstances, professional profiles, the sectors of the companies, and the companies' policies concerning remote or hybrid working modes. For example, employees of international tech firms were already in remote settings before COVID-19. They were used to online collaborations when COVID-19 hit, remaining remote during and after the lockdown. Nevertheless, they indicated the importance of in-person meetings for specific activities and project phases. Other innovation experts were employed in the heavy manufacturing sector, where physical presence is naturally more important, even though for the experts, physical presence was not an obligation since they held positions for which it was not required (in the management or innovation department). Experts from the fashion sector insisted on the importance of in-person meetings for creative sessions. Currently, their companies make physical presence obligatory. The obligation of 100 % in-person work was a rare case in our panel. The majority was fully remote during the lockdown and is currently working under different hybrid modes.

 https://doi.org/10.14512/tatup.33.2.49

The challenge these remote and hybrid modes set for managing the projects is to decide which activities are realized in which collaboration mode. Are there some meetings that require in-person attendance? Or is an always-remote option accepted? The hybrid mode creates several challenges for the participations during meetings (especially creative ones). The collaborators who are physically present tend to engage more with each other, whereas the colleagues connected online tend to be more passive than they are usually when engaging with the same group of colleagues on site. Several experts stated that the interruption culture is different when collaborating in hybrid modes and perceived this as linked to the absence of physical dimensions of inter-individual communication. To respond to the challenges of remote and hybrid working modes, some innovation managers stated that they require certain meetings to be in person, most often the kick-off meeting, especially when the colleagues do not know each other beforehand. For other innovation managers, physical presence remains always optional. Other types of meetings for which they partly insisted on physical presence were meetings with sensitive information, those concerning conflicts, or creative sessions.

The home office was perceived to increase efficiency by reducing travelling time and costs and providing more comfortable working settings. Other aspects were perceived as posing problems, for example by creating difficulties when onboarding new team members. Navigating the variety of different online collaboration tools (e.g., Microsoft Teams, Miro) and choosing the right modalities for each project phase was perceived as an essential management challenge for innovation projects. These circumstances led to considering the home office setup of the involved team members more thoroughly.

When asked about their personal preferences concerning their home office setup, many experts stated that ergonomic comfort is a key variable. A large screen (or several ones), eventually a modulable screen arm, a sit/stand desk, a comfortable chair, sufficient space, and a dedicated room were perceived as essential. These findings resonate with what I found in the current literature concerning ergonomic best practices in the home office (Gallagher et al. 2024). Remote and hybrid work environments have several impacts on the communication between colleagues and on the circulation of ideas during projects, letting new interpersonal and creative practices emerge. Several times, the experts expressed a certain nostalgia for physical collaborations, comparing for example online Miro sessions with a Whiteboard session in the office "Well, it's just not the same. You know, you can do anything on Miro, but you know, it's still not the same." (UX Researcher, public services).

Applied methods and creativity practices in online collaborations

The previous section described the current collaboration setup for innovation projects. To address the particularities of innovation projects, different methods and tools are applied. Some methods are the same as before COVID-19, while for others the communicational support has changed to be online. The most common methods stated by the experts are widely known innovation methods such as persona, design thinking, brainstorming, user journey, agile, stage gate, ease of use, Customer Effort Score, Net Promoter Score, and Problems-Concerns-Opportunities to name a few. Different customer acceptance tests are applied throughout the projects, as well as different objective and subjective measures. Some experts insist on the importance of physical check-in exercises for in-person, remote, and hybrid meetings (for example, breathing exercises, awareness techniques, or diverse mental representations exercises). Particularly for creative sessions, they underlined the importance of undisturbed attention of all participants to exploit the highest creative potential. This includes closing the laptop or putting the smartphone upside down. They primarily practice this not out of obligation, but by themselves as an example – leading by inspiration.

The innovation sector depends greatly on creativity (Paris 2008; Sawyer and Henriksen 2024). This is why there are specific positions for idea scouting: surveilling internal and external landscapes for ideas and filtering the most potential ones. But more than that, there are moments during innovation projects when creativity becomes crucial for each team member. Therefore, I asked what the innovation experts do to get, collectively and individually, into a creative state. Some stated that a creative mindset would be a permanent state they are in: "You know, I'm generally a creative guy, it's not like a different gear my brain gets in. I'm permanently thinking creative, so I don't have like any specific rituals or something" (Open Innovation Manager, automotive sector). Others do have specific habits when preparing for creative tasks. Examples of such routines are listening to music, walking around inside or outside the house, reading books, petting their cat, or having a dedicated, inspiring room for a creative mindset. Some explained that it depends on each situation and the associated emotions. Other rely on subconscious thinking. They familiarize themselves with the given task early on to internalize it, potentially leading to the famous Eureka moment in the shower or in bed – places experts also mentioned when it comes to creative moments. For others, a hyper focused tunnel vision mode is important, including turning off push notifications and email software. Several underlined the importance of a positive affective state. Positive affects were repeatedly stated as creativity enhancing, especially for interactions with colleagues.

Use case generation, variation, and anticipation

Use case generation is applied in innovation projects to define the concept precisely, to assure that the concept meets user needs, to communicate the concept to stakeholders, and to test it with the potential target group. Several use cases are generated, representing slightly different variations of the concept. Innovations are sometimes used in ways that were not intended by the involved teams. This can become a source of inspiration for new business opportunities. In other cases, it can lead to malevolent use cases. Therefore, innovation teams show an interest in anticipating different use case variations during an early project stage. An ap-

https://doi.org/10.14512/tatup.33.2.49

proach representing a general attitude of the innovation teams towards reinterpretations of the major use case is *'Celebrate the pivot!'* when an innovation concept undergoes an essential redefinition. This can be any use case variation becoming the major use case and does not necessarily represent a malevolent one. Other procedures to search for possible negative use cases is the identification of dark patterns and worst-case scenarios: "what would be the horror usage that can happen with that concept?" (Innovation Consultant, educational sector) Another procedure for identifying malevolent use cases is the definition of anti-personas. Certain innovation experts apply this procedure as an extension of the widely known persona method. It aims to identify anti-hero plots and negative use cases. This procedure is specifically dedicated to identifying malevolent use cases. During brainstorming sessions or group discussions, team members discuss which malevolent usages could occur with the innovation concept. This is the procedure aimed most explicitly at anticipating malevolent use cases according to our panel. Other common methods such as brainstorming or reverse brainstorming are used to identify possible usage extensions. Reverse brainstorming is used to turn positive use cases into negative ones. According to the innovation experts, this method aims at widening the list of generated ideas, not specifically at identifying possible negative use cases: "I think, again, it's, it's thinking around, you know, multiple scenarios and leveraging those scenarios to identify the key variables and now plotting those key variables from a negative angle and seeing what the potential impact could be." (Digital Innovation Manager Global, finance) Other experts test their concepts with a variety of user groups to observe possible usage extensions, aiming to create a large variety of possible use cases.

Discussion

Innovation projects are an economically important endeavor for companies, as they are indispensable for growth and staying competitive in saturated markets. I witness a high degree of ongoing differentiation and rationalizations during these projects, creating certain tensions when it comes to creative tasks (Habermas 1981). Consequently, the perspectives on this matter vary greatly according to different disciplines and personal preferences. I expected to find specific processes aimed at streamlining creativity during the use case generation and variation phase of innovation projects, for the anticipation of usage extensions, and especially for the identification of potentially malevolent use cases. This is not the case. In the interviews with the innovation experts, I could not identify specific cross-sectorial practices representing a higher degree of rationalization than for example the design thinking process. The same holds true for use case anticipation. I could not identify specific procedures used to anticipate malevolent cases. They were covered by practices such as describing an anti-persona. This was mainly done to search for new ideas and not specifically to avoid malevolent use cases. Other practices covering the aspect of use case anticipation were general creativity practices such as brainstorming or reverse brainstorming.

Surprisingly, according to the innovation experts, AI-based tools do not yet play a notable role in use case generation, variation, or anticipation. The innovation experts do use AI, but not specifically for the generation, variation, or anticipation of use cases. This is interesting because the currently available AI tools offer a high potential for these parts of innovation projects. Use case generation could be automated via AI narration tools (e.g., ChatGPT) in combination with visualization tools (e.g., Midjourney). Some experts stated that they could envision the integration of these tools for this task but that it is not yet done. Currently, these tasks are still completed manually by the relevant innovation teams. This point represents significant potential for the automation of innovation projects. Innovation ideas could develop into a multitude of different versions, becoming associated with corresponding use cases that get scripted and visualized in an automated, time and cost-effective manner. The absence is surprising considering the tremendous industrial potential for generative AI applications (Chui et al. 2023). Concerning the impact of AI on creativity see Dufour (2023), Grilli and Pedota (2024) and concerning its impact on an individual's well-being see Montag et al. (2023). Regarding the overall trend of process automation, I expected to find more specific AI-based procedures to generate use cases and use case variation in a more automated manner. It leaves room for discussion why this aspect of innovation projects has not yet seen a more profound impact by AI procedures. It is possible that this study was conducted too early, since AI narration and visualization tools have seen rapid progress during recent months. Indeed, the innovation experts did not describe notable applications of the Metaverse as a collaboration tool. This aspect is surprising considering the remote and hybrid nature of innovation projects, which could harness the potential of the Metaverse for online collaborations (Canivenc 2023). It should be expected that corporate innovation collaboration practices in particular (Sacépé

Surprisingly, according to the innovation experts, AI-based tools do not yet play a notable role in use case generation, variation, or anticipation.

 https://doi.org/10.14512/tatup.33.2.49

2019) will become more widely transferred to the virtual sphere. I expect such a tendency for the upcoming years, along with an increase in AI-based use case scenario development tools, keeping in mind that this might not be a particularity of innovation projects but an overall technological tendency. I assumed that AI assistants such as Microsoft's Copilot are spreading quickly and consequently would impact innovation projects and use case generation. It seems that it was too early for this study to find significant traces of such applications. It should be kept in mind that a rising integration of AI-based tools emphasizes the importance of screening for malevolent potential use cases. Since more aspects of use case generation and variation might become automated, the supervision of these scenarios might decrease. This becomes especially important for malevolent use cases.

Use case generation, variation, and anticipation represent the potential to direct possible usages of innovations. They imply the responsibility to promote or avoid certain use cases, including malevolent ones. This aspect seems underrepresented in the responses of our sample. The societal impact of innovation starts to get shaped during use case generation, even though the effective usages and societal implications of technologies remain difficult to anticipate (Rogers 1983). Regarding the growing societal importance of innovative technologies, the anticipation of use cases also becomes more important, especially regarding industrial and private AI applications (Mazzone and Elgammal 2019; Osei-Mensah et al. 2023). Usages guided by malevolent creativity with negative societal impacts are a specific case of usage extensions (Cropley et al. 2008). Current examples are the malevolent usages of the Metaverse, such as misinformation, financial crimes, or sexual harassment (Lepage and Mettrie 2022; Dwivedi et al. 2023). The procedures, methods, and tools applied during innovation projects, such as use cases, impact the final innovations, and in consequence the daily lives of users and their physical and affective dispositions (Martin-Juchat 2015), and therefore society at large. Regarding the growing societal impact of technological innovations, the different procedures involved in innovation projects are gaining research attention. In previous research, we analyzed the role of innovation methods and tools in general (Henke and Martin-Juchat 2021; Henke 2021), whereas this article focuses specifically on the role of use cases and the underlying communicational particularities.

The communicational structures underlying innovation practices have changed dramatically during the last years, greatly accelerated by the COVID-19 crisis. This article documents current practices as described by international innovation experts. The experts did not discuss cross-sectorial procedures for the anticipation of malevolent use cases. This may be due to the heterogeneity of the panel. The cross-sectorial design turned out as the major weakness of the applied methodology. The setup was chosen to identify general procedures, but the diversity of practices is too large. Either the quantity of conducted interviews should be increased, or the target group should be narrowed down. For the anticipation of malevolent use cases more specifically, I propose a research methodology focused on business-to-customer and product design innovation projects, where this phenomenon may be more easily visible.

Conclusion

I analyzed innovation practices based on interviews with cross-sectorial, international innovation experts (N = 14), asking about specific procedures that they use to generate, vary, and anticipate use cases. I indicated the current remote and hybrid nature of innovation projects – an aspect that can be seen not as a particularity of the innovation sector but as part of a broader technological tendency, accelerated via the COVID-19 crisis. This creates challenges for the management of innovation projects, especially for professional communication and for creativity sessions. Innovation practices vary greatly according to business-to-business vs. business-to-customer, product design vs. web design, and industrial vs. governmental projects. Anti-personas and worst-case scenarios are the procedures most specifically aimed at identifying malevolent use cases. Beyond them, the discourse of the experts did not enable the identification of cross-sectorial procedures for anticipating malevolent use cases. Generally, a multitude of use case variations is created during the process of use case generation, mostly aimed at widening the diversity and amount of concept ideas, and only in rare cases aimed specifically at identifying malevolent use cases. I could not identify a dedicated procedure that uses AI-based tools for use case generation, variation, and anticipation. This led to recognizing the unexploited potential of automating this part of innovation projects. Applying AI could allow innovation teams to generate, visualize, and evaluate a variety of different concept variations rapidly. The innovation experts noted the research potential of streamlining this idea towards an operational process.

The communicational structures underlying innovation practices have changed dramatically during the last years, greatly accelerated by the COVID-19 crisis.

The societal impact of innovation, especially technological innovation, is ever more increasing, necessitating further scien-

https://doi.org/10.14512/tatup.33.2.49

tific research to understand industrial practices and their societal consequences. The societal meaning of technologies depends not only their usage, but also the development process (Habermas 1968; Honneth and Sutterlüty 2012; Doerre et al. 2015). It therefore becomes important to understand the underlying industrial practices and preferences of the involved innovation experts. This article clarified the role of use case generation, variation, and anticipation in current innovation practices, adding complementary thoughts concerning AI and the Metaverse, which indicate promising research directions for upcoming studies.

Funding • This work received no external funding.
Competing interests • The authors declare no competing interests.

References

Albastaki, Fouzeya; Ubaid, Alaa; Rashid, Hamad; Alshamma, Ahmed (2022): Systematic literature review on the work from home concept. In: IEOM Society International (ed.): Proceedings of the 5th European international conference on industrial engineering and operations management. Rome: IEOM, pp. 733–747. https://doi.org/10.46254/EU05.20220147

Canivenc, Suzy (2023): Les espaces de travail virtuels. Au-delà du métavers. In: Repère – futures du travail 16, pp. 1–12. Available online at https://www.chairefit2.org/publications/les-espaces-de-travail-virtuels-au-dela-du-metavers/, last accessed on 03. 04. 2024.

Chui, Michael et al. (2023): The economic potential of generative AI. The next productivity frontier. In: McKinsey and Company, 14. 06. 2023. Available online at https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier#introduction, last accessed on 03. 04. 2024.

Cicek, Isik (2024): A systematic review of home-office-related studies in the business context during the COVID-19 pandemic term. In: Beatrice Orlando (ed.): Innovation capabilities and entrepreneurial opportunities of smart working. Hershey, PA: IGI Global, pp. 1–23. https://doi.org/10.4018/978-1-7998-8797-3.ch001

Cropley, David; Kaufman, James; Cropley, Arthur (2008): Malevolent creativity. A functional model of creativity in terrorism and crime. In: Creativity Research Journal 20 (2), pp. 105–115. https://doi.org/10.1080/10400410802059424

Doerre, Klaus; Lessenich, Stephan; Rosa, Hartmut (2015): Sociology, capitalism, critique. London: Verso.

Dufour, Hugues (2023): L'art face à l'IA. Vers un imaginaire augmenté. Paris: fyp.

Dwivedi, Yogesh et al. (2023): Exploring the darkverse. A multi-perspective analysis of the negative societal impacts of the metaverse. In: Information Systems Frontiers 25 (5), pp. 2071–2114. https://doi.org/10.1007/s10796-023-10400-x

Falahat, Mohammad; Chong, Shyue; Liew, Cindy (2024): Navigating new product development. Uncovering factors and overcoming challenges for success. In: Heliyon 10 (1), p. e23763. https://doi.org/10.1016/j.heliyon.2023.e23763

Gallagher, Kaitlin; Chambers, April; Lin, Jia-Hua; Jacobs, Karen; Robertson, Michelle (2023): Ergonomics advances in office work and work-from-home. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting 67 (1), pp. 346–349. https://doi.org/10.1177/21695067231192577

Grilli, Luca; Pedota, Mattia (2024): Creativity and artificial intelligence. A multilevel perspective. In: Creativity and Innovation Management, pp. 1–14. https://doi.org/10.1111/caim.12580

Guest, Greg; Bunce, Arwen; Johnson, Laura (2006): How many interviews are enough? An experiment with data saturation and variability. In: Field Methods 18 (1), pp. 59–82. https://doi.org/10.1177/1525822X05279903

Habermas, Jürgen (1968): Technik und Wissenschaft als "Ideologie"? In: Man and World 1 (4), pp. 483–523. https://doi.org/10.1007/BF01247043

Habermas, Jürgen (1981): Theorie des kommunikativen Handelns. Frankfurt a. M.: Suhrkamp.

Henke, Niklas (2021): Design methods and innovation projects as creativity consultancy. In: Revue française des sciences de l'information et de la communication 23, n.p. https://doi.org/10.4000/rfsic.11913

Henke, Niklas; Martin-Juchat, Fabienne (2021): The design turn for the management of public relations. Emerging challenges for communication professionals using design practices. In: ESSACHESS – Journal for Communication Studies 14 (2), pp. 153–170.

Honneth, Axel; Sutterlüty, Ferdinand (2012): Normative Paradoxien der Gegenwart. In: Hans-Georg Soeffner (ed.): Transnationale Vergesellschaftungen. Verhandlungen des 35. Kongresses der Deutschen Gesellschaft für Soziologie in Frankfurt am Main 2010. Wiesbaden: Springer, pp. 897–909. https://doi.org/10.1007/978-3-531-18971-0_80

Lepage, Xavier; Mettrie, Aymar de la (2022): Métavers. Enjeux, perspectives … et risques. In: Servir 515 (5), pp. 46–48. https://doi.org/10.3917/servir.515.0046

Martin-Juchat, Fabienne (2015): Capitalisme affectif. Enjeux et pratiques dans les organisations. In: Sylvie Alemanno (ed.): Communications organisationnelles, management et numérique. Paris: L'Harmattan, pp. 59–68.

Mazzone, Marian; Elgammal, Ahmed (2019): Art, creativity, and the potential of artificial intelligence. In: Arts 8 (1), p. 26. https://doi.org/10.3390/arts8010026

Montag, Christian; Nakov, Preslav; Ali, Raian (2023): Considering the IMPACT framework to understand the AI-well-being-complex from an interdisciplinary perspective. In: Telematics and Informatics Reports 13, p. 100 112. https://doi.org/10.2139/ssrn.4584349

OECD – Organisation for economic co-operation and development; Eurostat (2005): Oslo manual. Guidelines for collecting and interpreting innovation data. Paris: OECD Publishing. https://doi.org/10.1787/19900414

Osei-Mensah, Belinda; Asiamah, Elizabeth; Sackey, Rainbow (2023): Strategic communication and artificial intelligence. Reviewing emerging innovations and future directions. In: Archives of Business Research 11 (1), pp. 85–102. https://doi.org/10.14738/abr.111.13616

Paris, Thomas (2008): De Mozart à Pixar. Quand la créativité devient organisationnelle. In: Mode de recherche 10, pp. 3–15.

Rogers, Everett (1983): Diffusion of innovations. New York, NY: The Free Press.

Sacépé, Karine (2019): Le guide des espaces collaboratifs d'innovation. Paris: Institut open innovation.

Sawyer, Keith; Henriksen, Danah (2024): Explaining creativity. The science of human innovation. Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780197747537.001.0001

DR. NIKLAS HENKE
is scientific work package leader at Capgemini Engineering and part of *Future of People@Work*. His research focuses on creativity and the human body in the context of modern technological transformations. He teaches design and communication courses in Grenoble, Paris, Cologne, and Duesseldorf.