DELIVERABLE 2.1

Version 2.0

Design methodologies for addressing Ethical, Legal and Societal Aspects (ELSA) of military AI applications

Date	7 November 2024
Authors	Henning Lahmann, Bart Custers, Benjamyn I. Scott, Filippo Santoni de Sio, Marlijn Heijnen, Ivana Akrum
Version	Version 2.0

ELSA LAB DEFENCE

Abstract

This report, which is the first Deliverable in WP2 (D2.1) of the ELSA Lab Defence project, explores design methodologies to address ethical, legal and societal aspects (ELSA) of the use of military artificial intelligence (AI) applications. The methodologies mapped in this report serve as the starting point for developing within the project a comprehensive design methodology tailored to identifying ELSA issues in the use of AI technologies in the defence domain and providing guidance for designing military AI technologies that avoid or minimise ELSA issues. To map relevant technologies, a three-step approach is used.

First, to ensure the research results have usable benefits, the ELSA lab has developed three unique case studies. The use cases were selected as they generate a diverse range of ELSA-related problems, are pertinent for Dutch defence interests and, while relevant, have currently remained under-examined. The three established use cases are (1) Countering cognitive warfare using Early Warning Systems (EWS); (2) (non-lethal) autonomous robots; and (3) military decision-support systems.

Second, based on literature study, the most important ELSA issues regarding AI are investigated, without particular focus on the use of AI in the military domain. A total of nine values affected by the use of AI are identified and described (dignity, human agency and autonomy, responsibility, life and physical integrity, privacy and data protection, liberty, justice, democratic decision-making and political participation, and peace and international security). These values, as located in existing literature, are then put into the military AI context and linked to ELSA issues.

Third, existing ELSA design methods are identified and described. Most of these methods do not focus on defence and cannot directly be applied to the defence context, meaning that they may need to be adjusted and further tailored to military AI applications. A total of five major design methodologies are identified and described, all based on the concept of value sensitive design (VSD). These methodologies are: design for privacy/privacy by design (PbD), design for human agency and responsibility, including meaningful human control (MHC) and human oversight, explainable AI (XAI), and contestability by design, other approaches for design for human agency*, including cognitive and socio-cognitive engineering, human-machine teaming and team design pattern engineering, engaging society in the design of military systems, including Scandinavian participatory design, critical design, speculative design, social design, and (new) participatory design, and* design for emerging human-technology interactions, including coactive design and adaptive/adaptable automation. *These design methodologies are core design approaches and methods for addressing by design ELSA concerning new technologies. This provides an overview of the most relevant approaches and allows them to be applied to selected use cases.*

The introduction of new technology in defence offers opportunities, yet also creates risks. Introducing AI technology raises ethical, legal and societal issues. If AI is to be deployed and used responsibly, these and other aspects must constantly be considered in the design, implementation and maintenance of AI-based systems. Highlighting and understanding the different design methodologies from different sectors will allow for a holistic approach to be adopted, which can then be tailored to the specificities of defence and linked to the case studies.

Table of Contents

1. Introduction to ELSA in the Military

1.1 The ELSA Lab Defence

The suboptimal adoption of AI in defence organisations carries risks for the protection of the freedom, safety and security of society. Despite the vast opportunities that Artificial Intelligence (AI) technologies can present in the defence sector, there is also a variety of ethical, legal and societal (ELS) concerns. To ensure the successful use of AI technology by the military, ethical, legal and societal aspects (ELSA) need to be considered and their concerns continuously addressed at all levels. This includes ELSA considerations during the design, manufacturing and maintenance of AI-based systems, as well as its deployment and utilisation via appropriate military doctrine and training. This raises the question of how defence organisations can remain strategically competitive and at the edge of military innovation while respecting the values of citizens and applicable legal requirements.

This deliverable is part of the ELSA Lab Defence, which is a 4-year research project commissioned by the National Research Council in the Netherlands (NWO) on the ELSA of AI in defence.[1] The project aims to develop a future-proof, independent and consultative ecosystem for the responsible use of AI in the defence domain. In doing so, the ELSA Lab Defence will develop a methodology for context-dependent analysis, design and evaluation of ELSA of military AI-based applications. It builds upon and expands existing ethical design approaches like Value-Sensitive Design (henceforth also VSD), existing legal analyses of relevant principles and existing engineering approaches such as explainable AI and human-machine teaming. These methods are adapted to the specific defence context by conducting representative case studies, such as the use of (semi-)autonomous robots and AI-based methods against cognitive warfare. The lab also studies how defence personnel and society at large perceive the use of military AI, how this perception evolves over time, and how it changes in various contexts. Additionally, the ELSA Lab Defence monitors global technological, military and societal developments that could influence perception.

Although the focus of the research project is on different forms of AI in the defence sector, it focuses on three case studies (Chapter 2). One of these case studies, for example, is on unmanned aircraft.[2] The ELSA of unmanned aircraft in the defence domain typically concern issues like security (for people, objects, data or other aircraft), privacy (sensitive data, hindrance, annoyance, data collection or function creep), chilling effects, PlayStation mentality and post-traumatic stress disorder.[3] All of these play an important role in understanding the ELSA concerns, while highlighting

[1] https://elsalabdefence.nl/.
[2] Scott, B.I., Lahmann, H., Custers, B. (2023) Adopting AI in defense organisations requires further focus on ethical, legal and societal aspects, *ICUAS Magazine*, Vol. 1, Issue 2, p. 3-5.
[3] Custers, B.H.M. (2016) Flying to New Destinations: The Future of Drone Use, in: Custers, B.H.M. (ed.) The Future of Drone Use: Opportunities and Threats from Ethical and Legal Perspectives, Heidelberg: Springer/Asser Press.

the importance of utilising case studies that show nuances and the need for a context-specific approach.

The ELSA LAB Defence consists of several different work packages (WP), each with a set of deliverables. The substantial (i.e., non-managerial) WPs focus on developing an ELSA methodology for military applications (WP2), investigating perceptions and technological developments regarding AI in the military domain (WP3) and contextualisation and implementation of these findings, particularly in the context of several case studies (WP4). This report constitutes the first Deliverable (D2.1) in WP2, in which existing ELSA methodologies are identified and applied to the military domain.

1.2 The Aim of this Report

This report aims to explore design methodologies to address the ELSA issues caused by the introduction of AI in the military domain issues. To do that, it is, however, first necessary to obtain a clear understanding of these issues. While a lot of research already exists on the ELSA issues of new technologies, such as big data technologies, data science and AI,[4] and so-called AI ethics has become a growing field of research[5] most of this existing research is not applied or tailored to the military domain. Literature on specific ELSA issues raised by military AI exist, especially on so-called autonomous weapon systems[6].

The goal of this project is to develop an ELSA methodology for military AI applications outside the traditional scope. This methodology should be able to identify ELSA issues of AI technologies in the military domain and provide guidance for designing military AI technologies that avoid or minimise ELSA issues. In other words, this ELSA methodology can assist in identifying ELSA issues and in addressing these. An additional goal is to steer design to account for ELSA and to enable the evaluation of AI systems on ELSA. Developing such a methodology requires a clear understanding of ELSA issues, in this case, of military AI applications. This creates a chicken-and-egg problem: an overview of ELSA issues is required to develop a design methodology, but it is the ELSA methodology that provides this overview of the ELSA issues.

---

[4] See, for instance, La Fors, K., Custers, B.H.M., and Keymolen, E. (2019) Reassessing values for emerging big data technologies: integrating design-based and application-based approaches, *Ethics and Information Technology*, Volume 21, Number 3, p. 209-226. https://doi.org/10.1007/s10676-019-09503-4.

[5] See, for an introduction, Markus D. Dubber, Frank Pasquale, and Sunit Das, *The Oxford Handbook of Ethics of AI* (Oxford University Press, 2020); Mark Coeckelbergh, *AI Ethics*, The MIT Press Essential Knowledge Series (Cambridge, MA: The MIT Press, 2020).

[6] See, for instance, Nehal BHUTA et al., *Autonomous Weapons Systems : Law, Ethics, Policy* (Cambridge University Press, 2016), https://cadmus.eui.eu/handle/1814/43281; Robert Sparrow, 'Killer Robots', *Journal of Applied Philosophy* 24, no. 1 (2007): 62–77, https://doi.org/10.1111/j.1468-5930.2007.00346.x; Noel Sharkey, 'The Automation and Proliferation of Military Drones and the Protection of Civilians', *Law, Innovation and Technology* 3, no. 2 (1 December 2011): 229–40, https://doi.org/10.5235/175799611798204914; Stefan Buijsman et al., *Responsible AI in Military Applications*, Ethics and Information Technology, 2023, https://link.springer.com/collections/cghhijhhji.

This chicken-and-egg problem is addressed by taking an iterative approach to our project. This report constitutes the second cycle in this iterative approach (Deliverable 2.1).[7] Based on examining existing literature, the most important ELSA issues are identified (Chapter 3) [and it is discussed whether ELSA issues identified for domains other than the military are also relevant for the military via some examples]. Also, the most important existing ELSA design methodologies are examined (Chapter 3). Given that most of this literature is not applied or tailored to military AI applications, these approaches and methods are critically assessed in view of domain specificities (Chapter 4).

This is version 2.0 of Deliverable D2.1. Table 1.1 shows an overview of previous versions.

Version	Date	Major changes
1.0	2023	• First version with two cases studies, overview of affected values and potential value conflicts, and overview of existing design approaches and methods for ELSA methodologies.
2.0	2024	• A third case study was added. • Extension of relevant values and principles. • Revised structure to better link and express the values and principles with the different design approaches.

Table 1.1 Earlier versions of this deliverable D2.1.

The next steps of this iterative process can be found in subsequent reports of this research project. Deliverable 2.2 contains an ELSA impact assessment in military AI-based applications. Deliverable 2.3 describes ELSA design and development patterns for military AI-based applications. Deliverable 2.4 investigates algorithms for ELSA alignment in military AI-based operations. Deliverable 2.5 contains the final result: an ELSA methodology that can be deployed in military AI applications.

1.3 A Note on Methodology

The research in this report is based on desk research, mostly a literature study, complemented with online research. For mapping the most relevant values affected by AI (Chapter 3), the existing value lists of literature[8] on this topic and selected values that are most relevant when focusing on AI use in the military are combined. For mapping the most relevant design methodologies (Chapter 4), existing approaches were carefully examined mentioned in literature in ethics, law and social sciences. Through online searches, the list of design methodologies was verified and further expanded. In this

[7] Version 1.0 of Deliverable 2.1 can be found here: https://elsalabdefence.nl/wp-content/uploads/2024/06/Deliverable-2.1-ELSA-Defence_V1.0-1.pdf

[8] Custers, B., La Fors, K., Jóźwiak, M., Keymolen, E., Bachlechner, D., Friedewald, M., Aguzzi, S., (2017) Lists of Ethical, Legal, Societal and Economic Issues of Big Data Technologies (August 31, 2017). Report. Leiden: Leiden University., Available at SSRN: https://ssrn.com/abstract=3091018 or http://dx.doi.org/10.2139/ssrn.3091018

case no selection was made and all identified approaches are listed and described in this report. As a result, some design methodologies may overlap: some approaches were developed in response to earlier approaches and some were developed to elaborate on earlier approaches.

1.4 The Structure of this Report

Chapter 2 provides a description of the three case studies that the ELSA Lab Defence uses in the different deliverables. As deliverable 2.1 is introductory and foundational in nature, the case studies are described, alongside the relevant values and principles, and design methods.

Chapter 3 examines the most important ELSA issues regarding AI discussed in the existing literature about AI in general or AI applied to domains other than the military. As a first step towards the discussion of ELSA issues in the military domain, Chapter 3 provides examples of the values at stake and conflicting values when AI is applied in the military domain. This can also be part of a methodology to identify values that should be addressed given a specific human-AI system.

Chapter 4 provides an overview of the most important existing ELSA methodologies. Most of the literature (and the design methodologies described therein) are not applied or tailored to military AI applications. This means that it is not possible to simply pick and choose one of the existing ELSA methodologies and start applying it in the military domain. The methodologies first need to be adjusted and tailored to this specific domain. Additionally, a single methodology towards responsible (military) AI does not exist, not only because the applications of human-AI systems differ too greatly in terms of their goals, required tasks and capabilities, but also because the variety of contexts requires different methodologies. ~~Chapter 4 is structured within the AI lifecycle ranging from governance, design, development and operation based on the socio-technical feedback loop (SOTEF).~~

Chapter 4 also provides a first critical evaluation of these methodologies, to explore to what extent they are useful in the military domain and where and how they need to be modified. This evaluation is the final result of this report and the starting point for the next iteration, in which a more refined methodology specifically for military AI applications can be developed.

Chapter 5 contains the conclusions.

2. ELSA Lab Defence Use Cases

This chapter presents an analysis of three case studies, each offering unique insights into the topic. The selected case studies provide a diverse range of scenarios, highlighting different approaches, challenges and outcomes. By examining these 3 case studies, it will allow for the identification of key patterns, lessons learned, and best practices that can be applied to broader contexts. These case studies serve as a valuable resource for the remaining deliverables.

Three use cases of the (possible future) employment of AI-supported systems in the military domain were selected:

(1) countering cognitive warfare using Early Warning Systems (EWS)

(2) (non-lethal) autonomous robots[9]

(3) military decision-support systems

These use cases were selected because they represent autonomous AI systems, are already in use to some extent, generate a diverse range of ELSA-related problems, are highly relevant for Dutch defence, and have received relatively little attention in the research community so far. The first case is an example of a completely digital (i.e., virtual) AI system, whereas the second case is an example of a cyber-physical system. In military terms, the first example is non-kinetic, the second is kinetic, and the third showcases the use of an airborne system that requires human-machine cooperation.

Use case 1: countering cognitive warfare using Early Warning Systems

Cognitive warfare aims to change what people think, how they think and how they act. This military strategy has been fundamentally transformed by social media and AI (e.g. bots), which allow the (automated) spread of disinformation at an unprecedented scale, gradually weakening the recipients of the information and influencing public discourse around the globe. As one possible solution against these AI-driven adversarial tactics, it has been suggested to develop AI-enabled cognitive warfare monitoring and alert systems. Such an AI-based Early Warning System (EWS), which employs machine learning principles to parse vast amounts of text, image and video data that are being distributed and shared through social media or other online sources in real-time. The basic idea is to automatically detect coordinated adversarial behaviour that attempts to influence the target state's domestic audience by way of targeted, misleading information to achieve political outcomes. If the EWS is capable of issuing threat detection and warning notifications at the outset of such an adversarial information operation, then the authorities of the targeted state might be able to initiate certain countermeasures before the operation can start causing harm. For example, if an information operation tries to influence an upcoming democratic election by launching a large-scale disinformation campaign on social media, targeting ostensibly susceptible subgroups of the population, the EWS could alert authorities to enable them to cooperate with the social media

[9] See https://elsalabdefence.nl/use-cases/.

companies to either suppress the false and misleading messaging or to distribute effective counter-messaging to alleviate possible consequences.

Use case 2: (non-lethal) autonomous robots

The possibilities of fielding autonomously operating robots for military missions are manifold. In each scenario, one of the advantages of their deployment is that they could be tasked with carrying out missions in place of human soldiers and in this way save and preserve their lives, having removed them from the battlefield. Whereas weaponised robots that autonomously use force against humans have sparked a heated debate worldwide, many applications are conceivable that are far more realistic in the short term. For example, an autonomous, land-based robot could be equipped with sophisticated audio-visual and other sensors to survey a predefined area in enemy-held territory to collect information for ISR. Other possible scenarios involve autonomous minesweeping or securing a perimeter that human assets intend to occupy or pass through; a robot could also be used to engage an object without the risk of harm to humans present in the target area; or deployed for more critical tasks such as crowd control in a civilian-heavy urban environment using non-lethal weapons such as tasers or acoustic devices. To be sure, not only land-based autonomous robots are conceivable. There have been a number of military scenarios devised that revolve around the deployment of autonomous aerial or naval vehicles, and even sub-surface robots.

Use case 3: military decision-support system

Military decision-support systems are a class of software platforms that aim to assist military commanders with a vast array of critical mission control decisions on the tactical level, from the planning of an operation potentially all the way to concrete targeting decisions on the battlefield. The latest generation of such systems is usually built around so-called fusion architectures, i.e., virtual-physical infrastructures that integrate incoming data streams from various sources – such as sensors mounted on military assets in the field, satellites, surveillance unmanned aircraft, information gathered online, human and signals intelligence sources – and autonomously analyse the vast volumes of data to generate predictive recommendations for the human operators, who in theory at all times remain in control of the decision whether to follow these recommendations and to execute them through the chain of command. Battlefield management systems, as a sub-form of a military decision support system, strive to enhance command and control capabilities down to the individual military units in the field by enabling better situational awareness for

example by providing soldiers with tablet computers through which they can access relevant intelligence and the algorithmic recommendations at all times.

3. Relevant Values and Principles: Ethics and the Law

According to the Responsible Research and Innovation (RRI) approach, the design and deployment of emerging technologies should be built upon guidelines grounded in fundamental ethical and legal principles that have been critically and reflexively adopted by society. In line with the approach proposed by European ethical and policy documents such as the European Group on Ethics' (EGE) statement on robotics and autonomous systems,[10] the High-Level Expert Group on Artificial Intelligence's (AIHLEG) Guidelines for Trustworthy AI,[11] and the EU Expert Group report on ethical issues with driverless mobility,[12] the analysis and recommendations of this deliverable are guided by the ethical and legal principles described in this chapter.

The sources of these principles are laid down in the EU Treaties, the Charter of Fundamental Rights of the EU, and other applicable international law or stem from generally recognised ethical frameworks. Each of the sections in this chapter is structured as follows. First, each principle is described and defined by reference to its source. Next, it is assessed how the use of AI technologies may generally affect the respective principle. Finally, the implications for the employment of AI technologies in the military domain are evaluated. For that purpose, the different use cases put forward in Chapter 2 are considered.

3.1. Dignity

The first and most foundational value to be mentioned is dignity. Originating, in Western culture, from Enlightenment thought, the principle of human dignity generally demands that every individual has an inherent worth that translates to the right to be recognised and treated as a *person*. Crucially, this entails that individuals must at all times be treated as ends in themselves and not merely as a means to pursue an end.[13] Accordingly, the report of the EU Expert Group on Ethical Issues on Driverless Mobility states that "[e]very individual human possesses intrinsic worth that should not be violated or traded for the achievement of any other ends".[14] This elementary scope justifies the claim that human dignity is "the normative point of reference" of all other values and fundamental rights.[15] In ethics, dignity is the clearest way of taking the vulnerabilities of others into account. It is an expression of the human condition and protection of the human condition, which is inherently vulnerable.[16]

Aside from its philosophical grounding, dignity is explicitly recognised in various legal frameworks on national, regional and international levels. For instance, the German Constitution starts with guaranteeing the inviolability of human dignity (Article 1). The Constitution of South Africa proclaims

[10] European Group on Ethics in Science and New Technologies, 'Artificial Intelligence, Robotics and "Autonomous" Systems' (European Commission, 2018).
[11] European Commission, 'High-Level Expert Group on Artificial Intelligence ETHICS GUIDELINES FOR TRUSTWORTHY AI', 2019.
[12] Bonnefon, J.-F. et al. (2020) 'Ethics of Connected and Automated Vehicles: Recommendations on Road Safety, Privacy, Fairness, Explainability, and Responsibility', Independent Expert Report, 20 ff.
[13] Kumar, A. (2021) Kant on the Ground of Human Dignity. *Kantian Review*, Vol. 26, 435.
[14] Bonnefon et al.
[15] Id.
[16] Fineman, M.A. (2008) The Vulnerable Subject: Anchoring Equality in the Human Condition. *Yale Journal of Law & Feminism*, Vol. 20, No. 1, 2008.

that the republic is founded on the value of human dignity (Article 1(a)) and that "[e]veryone has inherent dignity and the right to have their dignity respected and protected" (Article 10). In its Preamble, the Indian Constitution affirms that the republic "assur[es] the dignity of the individual". On the European level, the Charter of Fundamental Rights of the EU sets out in its first operative article that "[h]uman dignity is inviolable. It must be respected and protected". The two central documents of international human rights law (IHRL), the Universal Declaration of Human Rights and the International Covenant on Civil and Political Rights (ICCPR), both recognise the critical importance of human dignity in its preamble and Article 1, respectively.

The principle of human dignity is furthermore recognised as a value protected by international humanitarian law (IHL), albeit only implicitly. Normatively, dignity is rooted in the principle of humanity, which, in inherent tension with the equally foundational principle of military necessity,[17] "underlie[s] and inform[s] the entire normative framework of international humanitarian law".[18] The International Criminal Tribunal for the former Yugoslavia (ICTY) observed in this context that "[t]he essence of the whole corpus of international humanitarian law as well as human rights law lies in the protection of the human dignity of every person. [...] The general *principle of respect for human dignity* is [...] the very *raison d'être* of international humanitarian law and human rights law".[19]

Legal frameworks commonly declare dignity 'inviolable'.[20] This implies that in its legal form, it is an absolute right, meaning that under no circumstances violations of this right are allowed. This sets the fundamental right to dignity structurally apart from other rights such as privacy, freedom of expression, or even the right to life, which can be infringed upon if certain overruling conditions are met. Due to this normative structure, debates in legal practice and scholarship usually focus on the question of what types of conduct fall within the protective scope of the right. Generally speaking, the concept of dignity is very broad but not clearly demarcated. Some cases are unambiguous. For example, torture is always a violation of human dignity.[21] That is irrespective of the issue whether torture in a particular instance aims at humiliating the victim; it is solely the fact that by definition, the purpose of such practice is the subjugation of the victim's ability to exercise their own free will, i.e., their autonomy. In the attempt to extract information against the person's will by means of torture, the person is treated purely as a means to an end.

Aside from this clear-cut case of a violation of dignity, many other instances are more controversial. This does not least hold true for the context of the employment of AI-supported applications. However, a connection can be drawn between dignity and the right to privacy. In the information society, the reputation of people is increasingly constituted by the data that are disclosed about them. Such disclosure of personal data can be voluntary or involuntary. As a result of this, people are increasingly judged upon their digital representation (the digital person) rather than as human beings of flesh and blood. [22] When an individual is no longer treated as a person with particular interests, feelings and commitments, but merely as a bundle of data, that person's dignity may be

---

[17] Corn, G.S. (2013) 'Humanity, Principle of', in R. Wolfrum (ed.), Max Planck Encyclopedia of Public International Law, para. 2.

[18] ICRC (2023) 'The Principles of Humanity and Necessity', <https://www.icrc.org/sites/default/files/wysiwyg/war-and-law/02_humanity_and_necessity-0.pdf>.

[19] *Prosecutor* v. *Furundžija*, IT–95–17/1–T, 10 December 1998, ICTY, para 183; on the status and protection of human dignity in IHL generally see Le Moli, G. (2021), Human Dignity in International Law, Chapter 4.

[20] See, as examples, Article 1 German constitution, and Article 1 Charter of Fundamental Rights of the EU.

[21] Luban, D. (2009). Human dignity, humiliation, and torture. *Kennedy Institute of Ethics Journal*, *19*(3), 211-230.

[22] Daniel J Solove, *The future of reputation: Gossip, rumor, and privacy on the interne*t (Yale University Press 2007).

compromised. The example of a machine-learning algorithm automatically classifying a Black couple as gorillas[23] demonstrates this problematic relationship as well. Generally speaking, practices like profiling can reinforce a tendency to regard persons as mere objects – that is, means to an end and not ends in themselves.[24] Therefore, interference with privacy in these instances implies interference with human dignity.

These issues are further exacerbated when it comes to the use of AI applications in the military domain. Here, one relevant debate grounded in the principle of dignity concerns the value of humanity and the risk of dehumanisation associated with new technologies. Some have argued, for instance, that any delegation of critical decision-making in the military – where lives are at risk – to automated or autonomous systems constitutes a violation of human dignity in itself,[25] and that Autonomous Weapon systems are a *mala in se*.[26] A further consideration relates to remote killing, e.g., via remotely operated unmanned aircraft. In such a scenario, two different aspects merit discussion. First, analogously to the above example of seeing persons as mere 'data bundles', often targeting decisions in these contexts will be made based on algorithmically generated recommendations by decision-support systems. Thus, a targeted individual will not be considered as a person with intrinsic worth but as an expression of aggregated data points. Second, when remote pilots deploy military unmanned aircraft for targeted killings, sometimes on the other side of the planet, it can appear to them that they are playing a game on a computer screen, whereas, in reality, they are piloting an aircraft that can kill people. This can arguably be seen as a violation of human dignity of the potential victims, as well as a moral injury for the operator.[27]

In more general terms, the increasing employment of AI in military systems, for example, to make targeting decisions, or more broadly in systems that carry out intelligence, surveillance and reconnaissance (ISR) missions, can be understood as furthering the trend of 'de-humanising' warfare, a development first described in the context of remotely controlled UAS.[28]

Depending on the larger context, the presence of autonomous robots used for military purposes within a civilian environment can have serious implications for the dignity of the resident population even if the machine itself is unarmed. At the height of the 'drone wars' against terrorist suspects in Afghanistan and Pakistan around the end of the first decade of the 21st century, a number of scholars observed increasing degrees of perpetual states of distress among civilians living in those areas that were frequently targeted by drone strikes. Constantly fearing for their lives, many people had begun abstaining from carrying out even the most trivial daily activities so as not to end up as accidental targets or 'collateral damage', with the mere sound or sight of a drone triggering severe states of

---

[23] Gray, R. (2015). Google apologizes after Photos app tags black couple as gorillas: Fault in image recognition software mislabeled picture. *The Daily Mail*.

[24] Lee A Bygrave, *Data protection law: Approaching its rationale, logic and limits*. *Information law series: Vol 10* (Kluwer Law International 2002).

[25] Asaro, P. (2012) 'On Banning Autonomous Weapon Systems: Human Rights, Automation, and the Dehumanization of Lethal Decision-Making', International Review of the Red Cross, Vol. 94, 687.

[26] Wallach, W. (2017) 'Toward a Ban on Lethal Autonomous Weapons: Surmounting the Obstacles', Communications of the ACM, <https://cacm.acm.org/opinion/toward-a-ban-on-lethal-autonomous-weapons/>.

[27] Coeckelbergh, M. (2013) 'Drones, Information Technology, and Distance: Mapping the Moral Epistemology of Remote Fighting', Ethics and Information Technology, Vol. 15, 87.

[28] Rogers and Holland Michel, Drone Warfare: Distant Targets and Remote Killings, in Romaniuk *et al*. (eds.), The Palgrave Encyclopedia of Global Security Studies (2020), <https://doi.org/10.1007/978-3-319-74336-3_33-1>.

anxiety.[29] With this in mind, it is not difficult to imagine how a roaming military robot in an urban setting might cause similar symptoms of distress despite not being equipped with weapons. The knowledge alone that whatever the robot observes may lead to lethal targeting decisions down the line might suffice to inhibit many otherwise normal human activities. If affected persons are, thus, barred from continuing to live their lives in the way they wish to, what is ultimately at stake is their dignity as human beings.

As indicated, it has been argued that virtually all the basic human rights (e.g., non-discrimination, freedom of expression, freedom of religion or privacy) and the core values in ethics (e.g., autonomy, non-maleficence, justice, or even privacy) are to some extent related to dignity. Given the specific relevance of the values of human agency and autonomy, which are particularly closely linked to dignity, these will be discussed separately in section 3.2.

3.2. Human Agency and Autonomy

The interconnected values of human agency and autonomy describe the concept that human beings should at all times be seen as free moral agents, who demand respect for the conditions of their agency. The idea that humans are to be conceived as autonomous, morally accountable agents stems, in Western culture, from Enlightenment thought, like the larger notion of dignity that it is derived from. Autonomy can be understood as "a person's effective capacity to act on the basis of beliefs, values, motivations, and reasons that are in some relevant sense their own".[30] Human agency, according to this conception, is one aspect of autonomy, describing more narrowly the person's effective ability "to enact decisions, make choices, and take charge of important aspects of their lives. This implies that they have meaningful options available to them and is thus contingent on a set of external requirements that need to be fulfilled".[31] It follows that autonomy has both an internal and an external dimension, as the concept encompasses the ability to form beliefs and formulate choices as well as the ability to act on them.

The general concept of autonomy is normatively derived from human dignity. Thus, on a legal level, it can be interpreted as falling within the protective scope of the *right* to dignity as laid down in positive legal frameworks (see above section 3.1). However, in the context of AI, it has been emphasised as a value that is critical enough to be considered in its own right.[32] To this end, many documents dealing with the challenges of AI have singled out autonomy as a core value to be protected going forward, for example the following: The Ethics Guidelines for Trustworthy AI drafted by the European Commission's High-Level Expert Group list "the principle for human autonomy" as one of its core ethical principles to ensure trustworthy AI systems. It states that "[h]umans interacting with AI systems must be able to keep full and effective self-determination over

[29] International Human Rights and Conflict Resolution Clinic, Stanford Law School and Global Justice Clinic, NYU School of Law, 'Living Under Drones: Death, Injury, and Trauma to Civilians from US Drone Practices in Pakistan' (2012) <https://chrgj.org/wp-content/uploads/2016/09/Living-Under-Drones.pdf>.
[30] Prunkl, C. (2024) 'Human Autonomy at Risk? An Analysis of the Challenges from AI', Minds and Machines, Vol. 34, 26.
[31] Id.
[32] See e.g. Laitinen, A. and Sahlgren, O. (2021) 'AI Systems and Respect for Human Autonomy', 4 Frontiers in Artificial Intelligence

themselves".[33] Similarly, the 2018 Declaration for a Responsible Development of Artificial Intelligence states that AI systems "must allow individuals to fulfil their own moral objectives and their conception of a life worth living".[34]

AI technologies can negatively affect both the internal and the external dimension of autonomy. For instance, a machine-learning algorithm might be deployed to manipulate a person's mental state so as to subliminally influence their behaviour.[35] Such practices are explicitly addressed by the EU AI Act, which prohibits any AI system that "deploys subliminal techniques beyond a person's consciousness or purposefully manipulative or deceptive techniques, with the objective, or the effect of materially distorting the behaviour of a person or a group of persons by appreciably impairing their ability to make an informed decision, thereby causing them to take a decision that they would not have otherwise taken in a manner that causes or is reasonably likely to cause that person, another person or group of persons significant harm".[36] For example, a toy equipped with an integrated voice assistant whose algorithms manipulate a child into acting recklessly would be prohibited under this rule.[37] The external dimension of autonomy, on the other hand, is primarily affected by increasing pressures to delegate tasks to AI systems, narrowing the scope of actual self-determination of human individuals.[38]

In the military domain, most debates merge the internal and external aspects under the encompassing notion of 'meaningful human control'. The question of how much control human individuals ought to have when deploying AI-supported military technology, and how much control they retain even if a system is nominally not fully autonomous, has been an ongoing academic and policy concern for well over a decade. Importantly, it transcends the more limited debate concerning the permissibility or desirability of genuinely autonomous weapon systems.[39] The general idea of 'meaningful human control' is that – no matter how intelligent and equipped with autonomous capabilities the respective military technology is – humans, not computers and their algorithms, should ultimately remain in control of, and thus morally responsible for, relevant decisions about (lethal) military operations.[40] This idea has achieved a wide consensus among scholars and

[33] EU Commission, Independent High-Level Expert Group on Artificial Intelligence, Ethics Guidelines for Trustworthy AI (2019), 12.
[34] Montréal Declaration for a Responsible Development of Artificial Intelligence (2018), 9.
[35] Prunkl, 25 f.
[36] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act), Article 5(1)(a).
[37] Casaburo, D. and Gugliotta, L. (2023) 'The EU AI Act Proposal(s): Manipulative and Exploitative AI Practices', KU Leuven Centre for IT & IP Law, <https://www.law.kuleuven.be/citip/blog/the-eu-ai-act-proposals-manipulative-and-exploitative-ai-practices/>.
[38] Prunkl, 25 f.
[39] This section is based on Filippo Santoni de Sio, *Human Freedom in the Age of AI*, First edition (New York: Routledge, Taylor & Francis Group, 2024), 149–77. For a philosophical discussion of the reader is referred to that chapter and to Filippo Santoni de Sio and Jeroen van den Hoven, 'Meaningful Human Control over Autonomous Systems: A Philosophical Account', *Frontiers in Robotics and AI* 5 (February 2018): 15, https://doi.org/10.3389/frobt.2018.00015.
[40] Article 36, 'Autonomous Weapons, Meaningful Human Control and the CCW' (Article 36, 2014), 36, http://www.article36.org/weapons-review/autonomous-weapons-meaningful-human-control-and-the-ccw/.

policymakers.[41] The requirement of Meaningful Human Control effectively captured that a mere human presence in the system, or the passive capacity to assess the behaviour of the system, would not be enough to protect human autonomous agency and to give human actors an effective capacity and opportunity to have the system behave according to their reasons and intentions and to remain responsible for each military action.[42] The ethical and political debates on the use of AI systems in the military are still ongoing.[43] Multiple definitions of 'meaningful human control' over AI-supported military systems have been proposed.[44]

The matter of control over military systems is not the only application context in which the question of autonomy and agency is relevant in the military context. Closer to the use cases envisaged by the EU AI Act and more specifically Article 5(1)(a) of the AI Act, any AI-supported system employed by a military to influence a target audience in any way likewise implicates the latter's autonomy. Thus, the conduct of 'cognitive warfare' tactics to exert effects on public opinion in an adversarial state by means of using machine-learning algorithms must be assessed against the values of autonomy and human agency.[45] The same consideration applies to countermeasures taken against such 'cognitive warfare', such as by deploying AI-based early warning systems or other such responses. Here, too, manipulative effects can occur that call for an evaluation of these values.

3.3. Responsibility

The value of responsibility is the counterpart of, and dependent on, human autonomy. Both individual persons and institutional stakeholders can and should be held morally and legally responsible for the consequences of their actions when it is appropriate to do so. At the same time, they should be given a fair capacity and opportunity to behave according to moral and legal expectations. Both these aspects are traditionally understood as requiring free will on the part of the moral agent,[46] i.e., *autonomy* as defined in section 3.2. Although the concept of free will has been challenged by developments in the neurosciences in recent decades,[47] the idea that human individuals generally freely decide to act in a certain way, are therefore *responsible* for their actions, and can thus be held *accountable*, is still the (implicit) foundation of criminal law as well as extensive parts of tort law.

---

[41] Sarah Knuckey, 'Governments Conclude First Debate on Autonomous Weapons Closes: What Happened and What's Next', 2014, https://www.justsecurity.org/10518/autonomous-weapons-intergovernmental-meeting/; Michael Horowitz and Paul Scharre, 'Meaningful Human Control in Weapon Systems: A Primer' (Center for a New American Security, March 2015), http://www.cnas.org/human-control-in-weapon-systems.

[42] Kerstin Vignard, 'The Weaponization of Increasingly Autonomous Technologies: Considering How Meaningful Human Control Might Move Discussion Forward', *UNIDIR Resources*, no. 2 (2014).

[43] Buijsman et al., *Responsible AI in Military Applications*.

[44] Merel Ekelhof, 'Moving Beyond Semantics on Autonomous Weapons: Meaningful Human Control in Operation', *Global Policy* 10, no. 3 (September 2019): 343–48, https://doi.org/10.1111/1758-5899.12665.

[45] For example: (1) in the form of large language models to generate false or misleading content to be distributed online to target audiences previously profiled as susceptible; or (2) in the form of AI-generated synthetic media ('deepfakes'))

[46] Talbert, M. (2024) 'Moral Responsibility', Stanford Encyclopedia of Philosophy, <https://plato.stanford.edu/archives/fall2024/entries/moral-responsibility/>.

[47] Focquaert, F., Glenn, A. and Raine, A. (2013) 'Free Will, Responsibility, and the Punishment of Criminals', in T.A. Nadelhoffer (ed.), *The Future of Punishment*, 247; Moore, J.G. (2016) 'Criminal Responsibility and Causal Determinism, Washington University Jurisprudence Review, Vol. 9.

The stability of this conception of (legal and moral) responsibility, however, has come under renewed pressure with the emergence of algorithmic 'agents' who appear to act autonomously although they are merely pieces of software. Proposed normative frameworks have responded to this predicament by emphasising the need to safeguard human responsibility when AI systems are being deployed in any context. For example, the EU AIHLEG's Ethics Guidelines for Trustworthy AI demand that "mechanisms be put in place to ensure responsibility and accountability for AI systems and their outcomes, both before and after their development, deployment and use".[48] The Montréal Declaration lists a "responsibility principle", proclaiming that "[o]nly human beings can be held responsible for decisions stemming from recommendations made by [AI systems], and the actions that proceed therefrom".[49]

This normative framing, however, is only the conceptual starting point of the problem of the 'responsibility gap' with 'learning automata', as presented by philosopher Andreas Matthias:[50] intelligent systems equipped with the ability to learn from the interaction with other agents and the environment will make human control over and prediction of their behaviour very difficult if not impossible. However, human responsibility requires knowledge and control. Therefore, as humanity faces a dilemma: either to go on with the design and use of learning systems, thereby giving up on the possibility of human persons being responsible for their behaviour, or preserve human responsibility, and thereby give up on the introduction of learning systems in society. Matthias' formulation of the responsibility gap has been quite influential, especially in relation to the development of autonomous weapon systems.[51]

This classic approach has been problematised in recent years. First, risks of gaps have been identified around not only the learning capacities of AI but, first and foremost, the opacity, complexity, and unpredictability that these systems generally display.[52] Second, it has been argued that while the development of AI systems may challenge our current moral and legal practices of attribution of responsibility,[53] this is not an insurmountable problem.[54] Third, an exclusive focus on the technical features of AI systems (e.g., complexity, opacity or 'autonomy' of learning algorithms) may be misleading. Responsibility gaps are due to a multiplicity of factors and are sometimes only aggravated by the presence of machines that learn and act on their own. For instance, sufficiently interconnected sociotechnical systems such as bureaucracies or corporates might also generate responsibility gaps. The infamous 'black box' problem due to the opacity and complexity of machine learning processes can also emerge due to the opacity or complexity of legal, organisational, or

[48] EU AIHLEG, 19.
[49] Montréal Declaration, 16.
[50] Andreas Matthias, 'The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata', *Ethics and Information Technology* 6, no. 3 (1 September 2004): 175–83, https://doi.org/10.1007/s10676-004-3422-1.
[51] Sparrow, 'Killer Robots'.
[52] Brent Daniel Mittelstadt et al., 'The Ethics of Algorithms: Mapping the Debate', *Big Data & Society* 3, no. 2 (December 2016): 1–21, https://doi.org/10.1177/2053951716679679.
[53] Ugo Pagallo, *The Laws of Robots: Crimes, Contracts, and Torts* (Springer, 2013), http://www.springer.com/gp/book/9789400765634; Ryan Calo, 'Robotics and the Lessons of Cyberlaw', *California Law Review*, no. Vol. 103, No. 3 (2015): 513–63, https://doi.org/10.2139/ssrn.2402972.
[54] Thomas W Simpson and Vincent C Müller, 'Just War and Robots' Killings', *The Philosophical Quarterly* 66, no. 263 (1 April 2016): 302–22, https://doi.org/10.1093/pq/pqv075.

social arrangements.[55] For instance, a military system used by a public organisation may be developed by a private company that does not want to share its data, or disclose its model or techniques. Fourth, while lawyers are well aware of the different meanings of the term 'responsibility',[56] the philosophical debate on 'the responsibility gap' has been criticised for not sufficiently taking into account these differences.[57] For a deeper look, 'the responsibility gap' identified in the classic philosophical debate[58] should rather be called a 'culpability' or 'liability' gap – the morally unsatisfying outcome that no human agent (or the wrong human agent) might be legally culpable or liable for harm caused by AI systems.

Other variations of the responsibility gaps are the 'accountability gap' and the 'active responsibility gap'. The accountability gap concerns the impossibility of having human persons and institutions *explain* what happened and comes in two forms: the public accountability gap, i.e., citizens not being able to get an explanation for decisions taken by public agencies, and a broader moral accountability gap – i.e., the reduction of human agents' capacity to make sense of and explain to each other the 'logic' of their behaviour.

Accountability gaps have been discussed widely in relation to the explainability of algorithms and AI systems generally.[59] In the military context, the debate was initially focused on autonomous weapon systems,[60] but has since expanded to address the use of AI systems by armed forces writ large.[61] Finally, while culpability and accountability gaps are primarily backwards-looking, that is focused on the possibility of calling people to respond for already occurred events, 'active responsibility gaps' are more forward-looking, that is concern people's capacity to act responsibly or 'take responsibility' for future events. In relation to AI, these gaps arise when persons designing, using, and interacting with AI are not sufficiently aware, capable, or motivated to see and act according to their moral obligations towards the behaviour of the systems they design, control, or use. In particular, they fail

---

[55] Guido Noto La Diega, 'Against the Dehumanisation of Decision-Making. Algorithmic Decisions at the Crossroads of Intellectual Property, Data Protection, and Freedom of Information', *Journal of Intellectual Property, Information Technology and Electronic Commerce Law* 9, no. 1 (31 May 2018), https://doi.org/10.31228/osf.io/s2jnk.

[56] Herbert Lionel Adolphus Hart, *Punishment and Responsibility* (New York and Toronto: Oxford University Press, 1968), http://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780199534777.001.0001/acprof-9780199534777; Joel Feinberg, *Doing & Deserving; Essays in the Theory of Responsibility,* (Princeton University Press, 1970).

[57] Filippo Santoni de Sio and Giulio Mecacci, 'Four Responsibility Gaps with Artificial Intelligence: Why They Matter and How to Address Them', *Philosophy & Technology* 34, no. 4 (14 December 2021): 1057–84, https://doi.org/10.1007/s13347-021-00450-x.

[58] Matthias, 'The Responsibility Gap'; Sparrow, 'Killer Robots'.

[59] Mittelstadt et al., 'The Ethics of Algorithms: Mapping the Debate'; Derek Doran, Sarah Schulz, and Tarek R. Besold, 'What Does Explainable AI Really Mean? A New Conceptualization of Perspectives', *CEUR Workshop Proceedings* 2071 (2 October 2017), https://arxiv.org/abs/1710.00794v1; Frank Pasquale, *The Black Box Society* (Harvard University Press, 2015).

[60] C Heyns, 'Report of the Special Rapporteur on Extra-Judicial, Summary or Arbitrary Executions', (United Nations, 2013); Chantal Meloni, 'State and Individual Responsibility for Targeted Killings by Drones', in *Drones and Responsibility: Legal, Philosophical and Socio-Technical Perspectives on Remotely Controlled Weapons*, by Ezio Di Nucci and Filippo Santoni de Sio (Routledge, 2016).

[61] Zhang, B. (2024) 'Accountability and Responsibility for AI-enabled Conduct', in R. Geiß and H. Lahmann (eds.), *Research Handbook on Warfare and Artificial Intelligence*, 216.

to ensure that these systems do not impact negatively on the rights and interests of other persons, but ideally, contribute positively to their wellbeing.[62]

3.4. Life and Physical Integrity

Together with dignity, the right to life belongs to the very core of fundamental rights as it constitutes the precondition for the enjoyment of all other rights.[63] It is explicitly acknowledged in all international human rights instruments and many national constitutional arrangements. Its foundational character is exemplified in the text of the ICCPR – the principal human rights treaty within the United Nations system – which states, in Article 6, that "[e]very human being has the *inherent* right to life" (emphasis added). Similarly, the ECHR provides in Article 2(1) that no one shall be deprived of his life intentionally. Even more succinctly, Article 2(1) of the Charter of Fundamental Rights of the EU determines that everyone has the right to life.

Although it has a less explicit place in the most important international human rights instruments, for the following reasons it is beyond doubt that physical or bodily integrity is also a fundamental value to be taken into account in the context of AI-supported systems. With its inherent connection to any human being's biological existence, it is closely related to the right to life. While it is explicitly enumerated in Article 3(1) of the Charter of Fundamental Rights of the EU, but neither in the ICCPR nor the ECHR, both these latter treaties are recognised to implicitly contain such a right. As such, physical integrity is said to tacitly underlie many of the provisions found in human rights law, being fundamental, for example, to the rights to security of the person, freedom from torture and cruel, inhuman and degrading treatment, or privacy.[64] The European Court of Human Rights (ECtHR) has held that the physical and moral integrity of the person falls under the right to respect for private and family life as established by Article 8(1) ECHR.[65] Some national jurisdictions explicitly recognise the value as a constitutional right as well, such as the Dutch (Article 11) or the German (Article 2(2)(1)) Constitutions. In common law jurisdictions, the right to bodily integrity is often subsumed under the principle of autonomy.[66]

The extensive scope of the right to life, while it ought not to be interpreted narrowly,[67] is not without limits. The ICCPR, for instance, provides that no one shall be 'arbitrarily' deprived of their life, which in principle implies that such deprivation by the state may be justified under certain circumstances, particularly pending the adherence to certain procedural safeguards; the most important and consistently mentioned are a valid basis in law, necessity, and proportionality of the lethal act of

[62] Santoni de Sio and Mecacci, 'Four Responsibility Gaps with Artificial Intelligence: Why They Matter and How to Address Them'.
[63] UN Human Rights Committee, General Comment No. 36: Right to Life, 3 September 2019, UN Doc. CCPR/C/GC/36, at para. 2.
[64] Report of the Special Rapporteur on the Implications for Human Rights of the Environmentally Sound Management and Disposal of Hazardous Substances and Wastes, 7 October 2019, UN Doc. A/74/48053, at para. 19.
[65] ECtHR, Case of X and Y v. The Netherlands, app. no. 8978/80, judgment, 26 March 1985, at para. 22.
[66] Herring, J. and Wall, J. (2017) 'The Nature and Significance of the Right to Bodily Integrity', Cambridge Law Journal, Vol. 76, 566.
[67] UN Human Rights Committee, General Comment No. 36: Right to Life, 3 September 2019, UN Doc. CCPR/C/GC/36, at para. 3.

state.[68] At the same time, while the situation of a public emergency, such as an armed conflict or an insurrection, allows a state to take measures derogating from its obligations under the ICCPR, this does not apply to the right to life under Article 4(2) of the ICCPR. The ECHR is explicit on the precondition of *absolut*e necessity and enumerates three lawful exceptions to the prohibition of deprivation of the right to life. It naturally follows that physical integrity is a value that is likewise not guaranteed absolutely. On the question of what constitutes a lawful interference with the right to physical integrity, such as being (1) in accordance with the law, (2) furthering a legitimate aim, and (3) necessary in a democratic society, there exists an extensive and nuanced body of jurisprudence by the ECtHR.[69]

In the context of AI-supported systems generally, the values of life and physical integrity are particularly relevant whenever such systems are used to assist or control physical systems that interact with human individuals, for example, with autonomous or semi-autonomous robots that are employed in health or elderly care.[70] Another important class of use cases concerns the safety of connected and autonomous vehicles (CAV) like self-driving cars[71] or UAS.[72] Here, the most important questions concern the establishment of adequate safety standards and other regulations that design organisations and manufacturers have to meet to ensure that their products do not imperil the life or limb of human beings, regardless of whether they are the users of the system or just otherwise affected individuals, such as pedestrians in an environment frequented by self-driving cars.[73]

In the military domain, the right to life and physical integrity are of great significance in the context of the deployment of AI-supported systems. At least four different dimensions should be considered to capture how the two values might be implicated.

First, there are conceivable scenarios in which the values of life and physical integrity play no role at all or at most a very insignificant one, for example, when AI is used to better manage maintenance cycles of military equipment.

As soon as such algorithms are connected to physical systems, a second dimension comes into play, which concerns general safety issues similar to those in civilian contexts as mentioned above. For instance, the execution of maintenance will involve factory workers, logistics personnel, or other individuals, who might be physically harmed or even killed if inadequate safety precautions cause AI-controlled mechanical parts to malfunction. Further, similar considerations apply to the case that military UAS navigate autonomously, or the armed forces develop maritime or ground vehicles that are capable of operating without human intervention. Such robots must be capable of reliably identifying obstacles, in particular any persons present in their path, and of acting accordingly so as to not imperil their health. UAS pose similar risks, for instance in the case that a UAS crashes into the ground or collides with another vehicle.

---

[68] UN Human Rights Committee, General Comment No. 36: Right to Life, 3 September 2019, UN Doc. CCPR/C/GC/36, at paras. 10–17.

[69] ECtHR, Guide on Article 8 of the European Convention on Human Rights, 31 August 2022, 10–14.

[70] Eduard Fosch-Villaronga and Tobias Mahler, Cybersecurity, Safety and Robots: Strengthening the Link Between Cybersecurity and Safety in the Context of Care Robots (2021) 41 Computer Law & Security Review 105528.

[71] Tom Michael Gasser, Fundamental and Special Legal Questions for Autonomous Vehicles, in Maurer *et al*. (eds.), Autonomous Driving (2016), 523.

[72] Benjamyn I. Scott (ed.), The Law of Unmanned Aircraft Systems, 2nd ed. 2022.

[73] BBC News, Uber in Fatal Crash Had Safety Flaws Say US Investigators, 6 November 2019, https://www.bbc.com/news/business-50312340.

A third dimension concerns AI-supported applications that aim at acting on adversaries without having the primary purpose of causing physical effects. For instance, in the future, militaries might run influence campaigns on enemy countries that utilise certain AI technologies such as deepfakes[74] or large language models to generate subversive messages.[75] Given the content of the messaging, it cannot be ruled out *prima facie* that the cognitive influence leads to harmful effects, for example, in the case that members from the target audience ingest poisonous substances due to disinformation.[76] Such consequences should be considered a possibility even if it remains very difficult to reliably establish a causal relationship between online disinformation and human attitudes and behaviour, not least in a legal sense.[77]

A fourth dimension of possible uses of AI-supported applications is of such a nature that the deprivation of life is the system's very purpose. This mainly concerns all those systems that are broadly employed to facilitate the conduct of armed operations by means of reconnaissance, intelligence collection, surveillance activities on the battlefield, or actual targeting and firing of weapons, either in a human-machine team setting or even autonomously. Generally speaking, in these scenarios the fundamental rights to life and physical integrity shift from a prohibition or protection by law to a guarantee of certain safeguards, principally but not exclusively to be found in applicable rules of international humanitarian law (IHL). As clarified by the UN Human Rights Council, lethal military operations in a situation of international or non-international armed conflict that comply with the cardinal rules of IHL such as the principles of distinction, proportionality, and precautions in attack, are normally not 'arbitrary' within the meaning of Article 6 ICCPR.[78] Another consideration in regard to life and physical integrity concerns not the person(s) targeted but any incidentally present bystanders, for instance, otherwise uninvolved civilians in an urban conflict environment.[79] Further, militaries might deploy AI-supported applications that do not have injury or deprivation of life as their primary purpose, but which can foreseeably cause this. Consider systems intended to affect objects only but that may nonetheless cause higher-level injury or death to persons, for instance by way of spreading disease or the destruction of infrastructure essential for the sustenance of the resident population.

A final consideration in the context of life and physical integrity are the risks to life and limb when a military decides *not* to use AI. This pertains to forgoing the opportunity to develop and deploy autonomous technologies in the field, which might mean that human members of the armed forces – like pilots or ground soldiers – have to come into harm's way in situations in which AI-supported

---

[74] Thom Waite, 'Digital Wildfire': How Deepfakes Became a New Frontier for Global Conflict, Dazed, 16 March 2023, https://www.dazeddigital.com/life-culture/article/58451/1/digital-wildfire-deepfakes-global-conflict-artificial-intelligence-socom-witness.

[75] Elias Groll, Researchers: Large Language Models Will Revolutionize Digital Propaganda Campaigns, Cyberscoop, 11 January 2023, https://cyberscoop.com/large-language-models-influence-operatio/.

[76] See Mostafa Shokoohi *et al.*, A Syndemic of Covid-19 and Methanol Poisoning in Iran: Time for Iran to Consider Alcohol Use as a Public Health Challenge? (2020) 87 Alcohol 25.

[77] Henning Lahmann, Infecting the Mind: Establishing Responsibility for Transboundary Disinformation (2022) 33 European Journal of International Law 411.

[78] UN Human Rights Committee, General Comment No. 36: Right to Life, 3 September 2019, UN Doc. CCPR/C/GC/36, at para 64.

[79] See Arthur Holland Michel, 'The Killer Algorithms Nobody's Talking About', Foreign Policy, 20 January 2020 <https://foreignpolicy.com/2020/01/20/ai-autonomous-weapons-artificial-intelligence-the-killer-algorithms-nobodys-talking-about/>.

technologies could take over. It has further been argued that AI-supported military decision-support systems could increase targeting precision and thus reduce the risks to civilians.[80]

3.5. Privacy and Data Protection

Privacy as a value is eminently critical in the context of AI-supported applications.[81] Originally, the right to privacy was conceived as primarily protecting private and family life. In the information age, it has gained importance in the guise of informational privacy, a concept that is closely related to the protection of personal data.[82] It describes the idea that individuals ought to have the right to know if, why, and how an entity such as a company or the state uses their personal data, and to be able to prevent such use unless justifying circumstances apply to a case at hand.

Privacy is widely established as a fundamental right, for instance in the UN Universal Declaration of Human Rights (Article 12), the ICCPR (Article 17), the EU Charter for Fundamental Rights (Article 7: privacy; Article 8: protection of personal data), and the ECHR (Article 8), as well as in many national legal frameworks. On the European level, the two fundamental pieces of legislation that concern the protection of data are the General Data Protection Regulation (GDPR) and the Law Enforcement Directive (LED). Neither framework applies to activities concerning national security, which includes most data processing conducted by an EU Member State's military. Whether the value of privacy is furthermore protected by IHL is subject to scholarly debate.[83]

Generally, there are two major ways in which the privacy of a person may be affected by developments in AI. Violations of privacy may occur:

(1) when AI tools process personal data, and
(2) when AI tools disclose privacy-sensitive patterns.

First, there is a connection between privacy concerns and the data that is used by AI tools. This has to do with the large amounts of data that AI systems are processing. Without these large amounts of data, machine learning-based AI tools cannot be trained and will deliver inadequate results. It is not hard to imagine that of all the data that AI tools process, some of it may be personal data (i.e., data relating to identified or identifiable natural persons). This is where personal data protection laws are relevant.

The GDPR establishes a set of principles and rules that data controllers need to take into account. All processing must be lawful, fair, and transparent.[84] Processing of personal data is lawful when the data subject has given consent, or when the processing of the data is necessary for the performance

[80] Meerveld HW and others, 'The Irresponsibility of Not Using AI in the Military' (2023) 25 Ethics and Information Technology 14.
[81] Omer, T., & Polonetsky, J. (2012). Privacy in the age of big data: a time for big decisions. *Stanford Law Review Online, 64,* 63.
[82] Custers, B.H.M., and Malgieri, G. (2022) Priceless data: why the EU fundamental right to data protection makes data ownership unsustainable, *Computer Law & Security Review,* Vol. 45, p. 1-13, https://doi.org/10.1016/j.clsr.2022.105683.
[83] Lubin, A. (2024) 'Big Data and the Future of Belligerency: Applying the Rights to Privacy and Data Protection to Wartime Artificial Intelligence', in: R. Geiß and H. Lahmann, Research Handbook on Warfare and Artificial Intelligence, 197.
[84] GDPR, Article 5(1).

of a contract, compliance with a legal obligation (usually for law enforcement purposes), or any of the other legal bases provided in Article 6 of the GDPR. This list of legal bases is exhaustive: when none of them apply, the collecting and processing of personal data is not allowed. The processing of sensitive data such as personal data revealing ethnicity, political or religious beliefs, genetic data, or data concerning sexual orientation is not allowed unless exceptions apply.[85] In case private actors process personal data, consent seems to be the most often used legal basis. There are many issues with how informed consent is given in the context of AI, for instance, because it may be difficult to explain and understand how the AI tools are processing personal data.[86] Further, power differences between the entity collecting and processing data on the one hand, and the data subject on the other, often make it impossible for the latter to refuse consent because the service in question is virtually indispensable.[87]

Second, privacy concerns arise not only from the utilisation of personal data but also from the outcomes of processing such data. This becomes particularly relevant in situations where the deployment of AI tools might unveil privacy-sensitive patterns. The potential for privacy issues becomes evident when AI tools predict information that individuals may prefer to keep undisclosed. In certain cases, data controllers may know more about data subjects than they do themselves, on life expectancy for example, the likelihood of their having serious illnesses or being in a car accident, risks of certain types of addiction, and estimates of well-being and happiness. For example, research shows that, based on only a few Facebook likes, it is possible to make a very accurate prediction of numerous sensitive personal attributes, such as sexual or political preferences.[88] Even when people do not or do not wish to reveal such aspects about themselves, it is possible to predict them based on other data that they (or others) have shared.[89]

The GDPR is not well attuned to the developments in AI.[90] Typically, AI tools benefit from processing large amounts of data, whereas the GDPR principles of collection and use limitation intend to restrict the collecting and processing of data. Also, AI tools offer the unique benefit of finding completely novel, unexpected patterns in data, something that can be very valuable in many contexts but at the same time does not match well with the purpose specification principle.

In the military domain, a lot of data that are processed are non-personal data that hardly affect privacy. For instance, surveillance data on landscapes, intelligence on a country's fleet, or flight routes of drones are unlikely to contain personal data. Other potential use cases, however, involve the processing of considerable amounts of personal data. One pertinent example is cognitive warfare. As identified from past instances of foreign influence campaigns – most prominently, the Russian targeting of British and American citizens ahead of the 2016 Brexit referendum and the US presidential elections – contemporary cognitive warfare campaigns often exploit personality traits as

---

[85] GDPR Article 9.

[86] Custers, B., Dechesne, F., Pieters, W., Schermer, B., and Hof, S. van der (2018) *Consent and Privacy*, in: Andreas Müller and Peter Schaber (eds.) Handbook of the Ethics of Consent, London: Routledge, p. 247-258.

[87] Bietti, E. (2020) 'Consent as a Free Pass: Platform Power and the Limits of the Informational Turn', Pace Law Review, Vol. 40, 307.

[88] Kosinski, M., Stillwell, D. & Graepel, T. (2012) Private traits and attributes are predictable from digital records of human behaviour, *Proceedings of the National Academy of Sciences* (PNAS), www.pnas.org/content/early/2013/03/06/1218772110.

[89] Custers, B.H.M. (2012) Predicting Data that People Refuse to Disclose; How Data Mining Predictions Challenge Informational Self-Determination, *Privacy Observatory Magazine*, Issue 3. See http://www.privacyobservatory.org/

[90] Zarsky T. (2017) Incompatible: the GDPR in the age of big data. Seton Hall Law Rev. 2017;47(4) Article 2.

derived from big data analysis of social media and other internet usage, harnessing such insights to micro-target individuals thus identified as susceptible to certain tailored political messages.[91] The detection and countering of such adversarial conduct by means of AI-supported applications such as an algorithmic early-warning system for cognitive warfare, might likewise require the collection and analysis of the state's own citizens' personal data to discover vulnerabilities and possible attack vectors. Furthermore, depending on the respective mode of operation, the algorithm's output might expose privacy-sensitive patterns of citizens' online habits. Therefore, privacy is one of the primary values to take into account when considering such an AI-based system.

Further, if an autonomous military robot is equipped with sensors to observe its environment, collect audio-visual data, and analyse the data about possible threats or targets by means of installed machine-learning algorithms, this has obvious privacy implications. This is particularly the case if the robot is fielded in urban surroundings in which a large number of civilians are present. Surveying their movements and quotidian activities quite obviously has considerable privacy implications. This concern is greatly exacerbated when the types of data collected include biometric information.[92] The right to privacy in both its input- and output dimension is affected if an AI-supported military decision-support system is used to detect and establish a target person's patterns of behaviour, for example for the purpose of predicting whether they are a terrorist or a member of a non-state armed group and thus subject to targeting decisions in armed conflict.[93]

3.6. Liberty

The value of personal liberty, understood in a narrower sense of not being detained or otherwise deprived of freedom of movement without justification, has been gaining increasing significance in the context of AI systems. As a human and civil right, it is explicitly mentioned in most relevant international legal instruments as well as in most national constitutions. In Article 9(1), the ICCPR stipulates that "[e]veryone has the right to liberty and security of person. No one shall be subjected to arbitrary arrest or detention. No one shall be deprived of his liberty except on such grounds and following such procedures as are established by law". The ECHR guarantees such a right in very similar terms in its Article 5, as well as Article 6 Charter of Fundamental Rights of the EU.

In civilian security and criminal justice contexts, AI is increasingly deployed for various tasks, most prominently to assess the risk of crime occurring in certain geographic areas or for an identified individual to become an offender ('predictive policing'),[94] which might lead to arrests and, thus, interfere with a suspected person's personal liberty. Other AI applications evaluate risks concerning

[91] Brian Resnick, 'Cambridge Analytica's "Psychographic Microtargeting": What's Bullshit and What's Legit' (*Vox*, 23 March 2018) <https://www.vox.com/science-and-health/2018/3/23/17152564/cambridge-analytica-psychographic-microtargeting-what> accessed 21 March 2023.

[92] Guo, E. and Noori, H. (2021) 'This Is the Real Story of the Afghan Biometric Databases Abandoned to the Taliban', MIT Technology Review, <https://www.technologyreview.com/2021/08/30/1033941/afghanistan-biometric-databases-us-military-40-data-points/>.

[93] Abraham, Y. (2024) '"Lavender": The AI Machine Directing Israel's Bombing Spree in Gaza', +972 Magazine, https://www.972mag.com/lavender-ai-israeli-army-gaza/.

[94] Meijer and Wessels, Predictive Policing: Review of Benefits and Drawbacks (2019) 42 International Journal of Public Administration 1031.

decisions concerning bail,[95] sentencing[96] or parole.[97] Decisions based on such algorithmic predictions directly affect the concerned individual's liberty as well.

AI systems that assist human operators with decisions concerning the deprivation of liberty of individuals will most likely gain more and more relevance in the military domain as well. Most importantly, in situations of armed conflict, military commanders might find themselves in the situation of having to detain either members of the opposing armed forces or civilians present in the theatre of conflict. International humanitarian law governs the possible grounds for lawful detention as well as procedural safeguards. AI might play a significant role when it comes to the detention of civilians on security grounds (Articles 41–43 and 78 of the Fourth Geneva Convention (GC IV)). Article 42 GC IV provides that such a measure may not be taken unless security considerations make it necessary. In procedural terms, IHL additionally guarantees the right to periodic review of the decision based on security-relevant information available at the time of the respective decision.[98] Aside from the rules of IHL, human rights law with its stricter safeguards for individuals generally remains applicable and governs questions not covered by IHL, such as those concerning the treatment of detained individuals or the right to a fair trial. In non-international armed conflicts, human rights law is the exclusive legal framework in this context.[99]

In contemporary military campaigns, civilian individuals are often detained not in the course of actual combat operations but as the result of intelligence-based risk assessments, carried out *ex ante*, that deem them a security threat.[100] More to the point, a military decision-support system might recommend taking a person who has been deemed a threat into preventative detention.[101] For example, Israeli security authorities claim to have used AI-enabled systems to parse social media and other online activity data of Palestinian teenagers in 2015 in order to stop an ongoing surge of "lone wolf" knife attacks by preventatively detaining persons who had engaged in suspicious activity on the internet.[102] Such conduct based on recommendations made by the decision-support system has a direct impact on the liberty of the concerned persons.

---

[95] Simonite, Algorithms Were Supposed to Fix the Bail System. They Haven't, Wired, 19 February 2020, https://www.wired.com/story/algorithms-supposed-fix-bail-systemthey-havent/.

[96] Donohue, A Replacement for Justicia's Scales? Machine Learning's Role in Sentencing (2019) Harvard Journal of Law & Technology 657.

[97] Singh *et al*., Predicting Parole Hearing Result Using Machine Learning, 2017 International Conference on Emerging Trends in Computing and Communication Technologies, https://ieeexplore.ieee.org/document/8280342.

[98] International Committee of the Red Cross, Procedural Principles and Safeguards for Internment/Administrative Detention in Armed Conflict and Other Situations of Violence, 2005.

[99] Pejic, Procedural Principles and Safeguards for Internment/Administrative Detention in Armed Conflict and Other Situations of Armed Violence (2005) 87 International Review of the Red Cross 375, 378.

[100] Goodman, R. (2009) 'The Detentions of Civilians in Armed Conflict', The American Journal of International Law, Vol. 103, 48.

[101] Ashley Deeks, 'Detaining by Algorithm', Humanitarian Law & Policy Blog, 25 March 2019, <https://blogs.icrc.org/law-and-policy/2019/03/25/detaining-by-algorithm/>.

[102] Amos Harel, 'How Israel Stopped a Third Palestinian Intifada', Haaretz, 4 October 2019, <https://www.haaretz.com/israel-news/2019-10-04/ty-article/.premium/how-israel-stopped-a-thirdpalestinian-intifada/0000017f-e355-df7c-a5ff-e37f99d30000>.

3.7. Justice

Justice is a further core value of immense significance in the context of the development and deployment of AI-supported technologies. As a fundamental ethical principle, in its broadest sense justice can be understood as seeking to ensure that every person is treated in a way that is equitable and fair.[103] In a legal context, it is often enshrined in rules that guarantee equality before the law and non-discrimination (see e.g., Articles 20, 21 and 23 Charter of Fundamental Rights of the EU).

To this end, in the context of AI technologies, the Montréal Declaration holds that AI systems "must be designed and trained so as not to create, reinforce, or reproduce discrimination based on – among other things – social, sexual, ethnic, cultural, or religious differences".[104] The EU High-Level Expert Group's Ethics Guidelines for Trustworthy AI state that "[t]he development, deployment and use of AI systems must be fair. [...] The substantive dimension implies a commitment too: ensuring equal and just distribution of both benefits and costs and ensuring that individuals and groups are free from unfair bias, discrimination and stigmatisation".[105] Similarly, justice "concerns the question of how we ought to distribute fairly the benefits and burdens of newly emerging technologies. Injustice occurs when the benefits to which an individual is entitled are unjustifiably denied, or when some burden is unduly imposed upon somebody without adequate compensation".[106]

How the development and deployment of AI technologies can negatively impact the principle of justice is highly context dependent. For example, the Independent Expert Report on the Ethics of Connected and Autonomous Vehicles explains that "CAVs should provide equality of access to mobility for all and should be calibrated by developers to reduce disparities in exposure to harm between categories of road users".[107] Over the past years, many cases of algorithmic decision-making leading to discriminatory outcomes have surfaced, for example in the field of HR recruitment[108] or government benefits.[109] In the security realm, the use of AI-supported tools to predict the occurrence of crime in certain neighbourhoods or by certain persons[110] or the likelihood of recidivism in criminal proceedings[111] has also led to proven discrimination against individuals from minority groups. Another, broader, concern related to justice is that of the shift of power that AI and digital technologies have created toward unaccountable actors like big tech companies, and the

---

[103] David Miller, 'Justice', in Edward N. Zalta and Uri Nodelman, *The Stanford Encyclopedia of Philosophy* (Fall 2023 ed.), <https://plato.stanford.edu/archives/fall2023/entries/justice/>.
[104] Montréal Declaration, 13.
[105] High-Level Expert Group, 12.
[106] Bonnefon, J.-F. et al. (2020) 'Ethics of Connected and Automated Vehicles: Recommendations on Road Safety, Privacy, Fairness, Explainability, and Responsibility', Independent Expert Report, 24.
[107] Id.
[108] Alina Köchling and Marius Claus Wehner, 'Discriminated by an Algorithm: A Systematic Review of Discrimination and Fairness by Algorithmic Decision-Making in the Context of HR Recruitment and HR Development' (2020) 13 Business Research 795.
[109] See e.g., Melissa Heikkilä, 'Dutch Scandal Serves as a Warning for Europe Over Risks of Using Algorithms', Politico, 29 March 2022, <https://www.politico.eu/article/dutch-scandal-serves-as-a-warning-for-europe-over-risks-of-using-algorithms/>.
[110] Will Douglas Heaven, 'Predictive Policing Algorithms Are Racist. They Need to Be Dismantled' (2020) MIT Technology Review, <https://www.technologyreview.com/2020/07/17/1005396/predictive-policing-algorithms-racist-dismantled-machine-learning-bias-criminal-justice/>.
[111] Brianna Rauenzahn, 'Addressing an Algorithmic PATTERN of Bias' (2022) The Regulatory Review, <https://www.theregreview.org/2022/05/10/rauenzahn-addressing-an-algorithmic-pattern-of-bias/>.

corresponding capacity for domination and exploitation of other societal actors, especially vulnerable ones.

As in civilian use contexts, the way AI systems employed by military organisations could foreseeably infringe on the principle of justice in many different ways. For example, if employed for the purpose of security detention during armed conflict, as explained above, training datasets that include discriminating or otherwise unfairly biased source data will lead to unjust results analogous to those in the criminal justice context. The same applies to the deployment of algorithmic decision-support systems for the purpose of ISR or targeting. Here, too, biased data may lead to, for example, erroneously designating a person as a terrorist or otherwise irregular militant that may be engaged by force.[112]

3.8. Democratic Decision-Making and Political Participation

Democracy and political participation are regulated in a series of individual and collective rights and represent important values that are to be accounted for when certain AI applications are considered by militaries. From a legal perspective, the values consist of and are guaranteed by a bundle of human rights that seek to ensure the functioning of democratic processes in liberal societies. The most important individual rights in this context are the right to freedom of expression and the right to freedom of information per Article 19 ICCPR, Article 10 ECHR and Article 11 Charter of Fundamental Rights of the EU; the freedoms of assembly and association per Article 21 ICCPR, Article 11 ECHR and Article 12 Charter of Fundamental Rights of the EU; as well as the right to vote and to be elected, which is guaranteed by Article 25 of the ICCPR. In their collective manifestation, these values find their expression in the right to self-determination in Article 1 of the ICCPR.

Despite being mentioned less frequently in the context of the deployment of AI systems in the military, these values might gain relevance concerning certain applications. If militaries use AI-generated textual or audio-visual content to influence operations against adversarial target populations, such activities may well interfere with the ability to free and uninhibited democratic decision-making, for example, if the conduct distorts the information environment that is necessary for the orderly execution of free and fair elections. Whereas influence operations may only in very limited circumstances be prohibited by international humanitarian law,[113] they nonetheless have important implications for these democratic values in general.[114]

The values of democratic decision-making and political participation may likewise be affected if an AI-supported monitoring and alert system is deployed to counter the threat of cognitive warfare. At its core, a system designed to counter cognitive warfare works with the baseline assumption that there exists potentially harmful information that the deploying state's civilian population must be protected from. In a liberal-democratic society built around foundational principles such as freedom of information and freedom of expression, this is a precarious proposition. As the former UN Special

---

[112] Kathleen McKendrick, 'Artificial Intelligence Prediction and Counterterrorism' (2019), Chatham House, <https://www.chathamhouse.org/sites/default/files/2019-08-07-AICounterterrorism.pdf>.

[113] Lahmann, Protecting the Global Information Space in Times of Armed Conflict (2020) 102 International Review of the Red Cross 1227.

[114] Ohlin, Did Russian Cyber Interference in the 2016 Election Violate International Law? (2017) 95 Texas Law Review 1579; Lahmann, Information Operations and the Question of Illegitimate Interference Under International Law (2020) 53 Israel Law Review 189.

Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression pointed out concerning the closely related concept of disinformation, it is an "extraordinarily elusive concept to define in law, susceptible to providing executive authorities with excessive discretion to determine what is disinformation, what is a mistake, what is truth".[115] The freedom of expression in principle encompasses a right to discuss and even disseminate information "even if it is strongly suspected that this information might not be truthful".[116] In his foundational 1859 essay *On Liberty*, JS Mill's famously argued that also false information and opinions that are widely considered wrong should be left free to circulate: because they may turn out to be (partially) true or correct in the future (knowledge evolves and values changes), and because, even if they remain false or wrong, they will force people to explain to others and remind themselves why they are so.[117] To some extent, false information and wrong opinions may contribute to keep a liberal democracy healthy, by supporting a (self-)critical attitude and a capacity to engage with opposing views and values. It follows that citizens must be free to choose their sources of information no matter whether such sources seek to influence them on behalf of a foreign power. It is part of the essence of liberal democracies that they are capable of accommodating a diverse and heterodox media ecosystem that comprises publications and other information sources of doubtful provenance. Interfering with this freedom would thus not simply infringe upon the individual civil rights of citizens but arguably compromise the process of democratic decision-making itself. It can be interjected that an AI-based system that merely acts as a detection and early warning instrument does not by itself impact any of these values; however, it seems safe to assume that once a foreign influence campaign has been flagged by the system, steps to counter such conduct will be initiated, and it would be artificial to isolate the necessary first step from this larger context. Furthermore, within an environment as inherently complex as the online information ecosystem, a considerable number of false positives are to be expected, which further imperils these values.

It has further been argued that the extended availability of AI and data-driven technologies in warfare may increase the reliance of governments on these systems for their decision and action, thereby reducing the role of political deliberation, i.e., leading to an increased 'depoliticization' of warfare.[118]

3.9. Peace and International Security

Peace and international security are crucial values at the heart of many considerations regarding the use of AI by armed forces. In its legal framing, they underpin the Charter of the United Nations, most prominently in the organisation's restatement of its principal purpose to "maintain international peace and security".[119] They, thus, constitute the foundational principle of the post-World War II global order. Although the precise content of the values in their legal iteration remains contentious and not well defined, at their minimum they prescribe the absence of armed hostilities between states.

[115] UN Doc. A/HRC/44/49, 23 April 2020, at para. 42.
[116] ECtHR, Salov v. Ukraine, 2005, at para. 113.
[117] John Stuart Mill, *On Liberty*, ed. Gertrude Himmelfarb, Reprinted, Penguin Classics (London: Penguin Books, 1985).
[118] Santoni de Sio, *Human Freedom in the Age of AI*, 105–6.
[119] UN Charter, Article 1(1).

The use of AI systems in military contexts has implications for the values of peace and international security in at least two ways.

First, the increasing employment of AI in weapon systems, for example, to make targeting decisions, or more broadly in systems that carry out ISR missions, can be understood as furthering the trend of 'de-humanising' warfare, a development first described in the context of remotely controlled armed UAS.[120] As this type of military operation vastly reduces the risk for military personnel, various scholars have argued that the trend has led to a creeping dissolution of the limits of warfare, in turn rendering perpetuated, low-intensity armed conflict without clearly defined goals much more likely.[121] This development, which is expected to become more entrenched as more states develop AI technologies for military applications, might have seriously destabilising effects on peace and international security in the mid- to long-term.

Second, a different class of AI systems might be used in the future to make or facilitate decisions to use armed force against another state, for instance, by algorithmically analysing vast amounts of intelligence information on the activities of adversaries, such as troop movements or arms build-up.[122] Variations of such systems are already in use given the massive quantities of data that contemporary intelligence agencies gather through automated electronic means every day and that cannot possibly be parsed by human operators. Under the existing *jus contra bellum* established by the prohibition of the use of force pertaining to Article 2(4) of the UN Charter, states are only permitted to use military force against another state in a situation of self-defence in response to an armed attack per Article 51 of the UN Charter or with a mandate by the UN Security Council. If AI is employed to assist with determining whether such an armed attack has begun or is imminent, this has wide-ranging and potentially momentous consequences for the maintenance of peace and international security and, thus, for the existing global order as established by the United Nations.[123] In 1983, the Soviet Union official Petrov allegedly saved the world from a nuclear catastrophe by overruling – in breach of his duties – the decision of an automated system who had wrongly identified an attack from the US and was activating a military response[124]. This story is often used as a reminder of the risks of automated systems for peace and security.

---

[120] Rogers and Holland Michel, Drone Warfare: Distant Targets and Remote Killings, in Romaniuk *et al*. (eds.), The Palgrave Encyclopedia of Global Security Studies (2020), <https://doi.org/10.1007/978-3-319-74336-3_33-1>.

[121] Bhuta and Mignot-Mahdavi, Dangerous Proportions: Means and Ends in Non-Finite War (2021) Asser Research Paper 2021-01; Lahmann, The Future Digital Battlefield and Challenges for Humanitarian Protection: A Primer, Geneva Academy Working Papers, April 2022, 24.

[122] McChrystal, AI Has Entered the Situation Room, Foreign Policy, 19 June 2023, <https://foreignpolicy.com/2023/06/19/ai-artificial-intelligence-national-security-foreign-policy-threats-prediction/>.

[123] Deeks, Lubell and Murray, 'Machine Learning, Artificial Intelligence, and the Use of Force by States' (2019) 10 Journal of National Security Law & Policy 1; Dominik Steiger, 'Employment of AI in Decisions on the Use of Force', in: Robin Geiss and Henning Lahmann (eds.), Research Handbook on Warfare and Artificial Intelligence, 2024.

[124] Dylan Matthews, '41 Years Ago Today, One Man Saved Us from World-Ending Nuclear War', *Vox*, 26 September 2018, https://www.vox.com/2018/9/26/17905796/nuclear-war-1983-stanislav-petrov-soviet-union.

4. Design for Ethical, Legal and Societal Values in Defence

This chapter reviews existing generic methodologies to identify and address ethical, legal and societal values in the development and design of new technologies, examines existing attempts to apply these to the development of AI, and critically assesses the promises and limitations of these methodologies in relation to AI in the defence domain.

This chapter is structured through the lens of Value-Sensitive Design (VSD), which is one of the first attempts at establishing a methodology to systematically design ethical values into (digital) technologies. It is one of the predominant approaches to design for values[125]. It is an overarching methodology, under which many other methodologies can be grouped. Not much literature exists on reporting on the actual implementation of VSD methodologies in military technology development[126]. This can be due to a lack of this implementation but also to a lack of reporting in the academic literature. This chapter fills this gap by investigating VSD in non-military contexts and how this could be applied in military context.

The methodologies in this chapter are presented as implementations of or in relation to VSD. This chapter does not give complete accounts of the history behind each methodology and acknowledges that not all methodologies were created with VSD in mind. By presenting the methodologies from the perspective of VSD, this chapter brings a first type of structuring to the methodologies. The methodologies relate to the values identified in chapter 3.

Section 4.1 and 4.2 introduce the concepts of 'design' and Value-Sensitive Design respectively. Section 4.3 focuses on VSD in defence. Section 4.4 outlines existing approaches for implementing VSD in non-military contexts that may be used as a starting point for a military implementation ('design for privacy'). Section 4.5 presents some existing theoretical proposals to use VSD to design for specific ethical, legal and societal values in the military ('design for human agency and responsibility). Section 4.6 connects this proposal with existing engineering and design methodology ('cognitive and socio-cognitive engineering'; human-machine teaming). Section 4.7 presents some avenues for future research to apply alternative design approaches (Critical, speculative, social, new participatory design) to promote a more critical societal discussion and to give people more power in the design of military systems. Section 4.8 presents some existing engineering approaches to design for emerging values and interactions.

4.1. Expanding the Concept of 'Design'

In general, the AI-lifecycle is considered to comprise various stages, ranging from problem definition and listing requirements to development and deployment. Although the concept of design is mostly used to describe the AI-lifecycle stage in which user requirements translate to specific AI

[125] Jeroen Van Den Hoven, Pieter E. Vermaas, and Ibo van de Poel, *Handbook of Ethics, Values, and Technological Design: Sources, Theory, Values and Application Domains* (Dordrecht: Springer, 2015), https://doi.org/10.1007/978-94-007-6970-0.
[126] Christine Boshuijzen-van Burken, 'Value Sensitive Design for Autonomous Weapon Systems – a Primer', *Ethics and Inf. Technol.* 25, no. 1 (11 February 2023), https://doi.org/10.1007/s10676-023-09687-w.

functionalities, it can also be used in a broader sense. The concept of design can extend towards the context, human interaction or organisation in which the AI system is used instead of only focusing on the AI system itself. This is a very important consideration about AI systems that keep on 'learning', adapting and evolving after they are first released. This means that the 'design' of an AI-system is never really finished, many actors other than the designers strictly conceived contribute to their development, and all aspects determining the effects of an AI-system should be considered even long after the AI-system was built.

4.2. Design for Values: Value-Sensitive Design (VSD)

According to VSD, human values should be explicitly and proactively accounted for right from the early stages of and throughout the technology design process.[127] To do so, a multidisciplinary effort is required and a methodology that includes three elements: a conceptual, an empirical, and a technical analysis.

The conceptual analysis consists of the theoretical study of the ethical values potentially involved in the technology: what do we want to achieve with this technology (e.g., efficiency, well-being and justice) and how do we want to achieve it (e.g., through designing for safety, environmental sustainability). It also requires an analysis of the stakeholders potentially affected by it: 'primary stakeholders' (actors who will directly interact with the technology, typically as users) and 'secondary' stakeholders (actors who will be affected by the technology indirectly. For example, in the case of an automobile, not only the drivers and passengers, but also other road users and the environment.

The empirical analysis consists of an empirical study of the stakeholders' values, attitudes, and thinking concerning the specific technology and the specific domain of its application. This empirical analysis includes detailed investigation of what the relevant stakeholders think and value about a certain technology and what their needs, hopes, fears, and attitudes are.

The technical investigation addresses how to translate the specific values and requirements identified in the conceptual and empirical analyses into the technical features of the system. The VSD process must be iterative and circular as it does not flow only once and in only one direction. The conceptual investigation influences the empirical and technical research by identifying and defining relevant values, but at the same time, abstract concepts are also integrated by stakeholders' interpretations of values in context and by technical consideration about the functioning of the relevant systems. In addition, design solutions and prototypes should be conceptually and empirically analysed and assessed, which may lead to new cycles of analyses at all three levels and so on.[128]

---

[127] Batya Friedman, 'Value-Sensitive Design', *Interactions* 3, no. 6 (1996): 16–23, https://doi.org/10.1145/242485.242493; Janet Davis and Lisa P Nathan, 'Value Sensitive Design: Applications, Adaptations, and Critiques', in *Handbook of Ethics, Values, and Technological Design: Sources, Theory, Values and Application Domains*, ed. Jeroen van den Hoven, Pieter E Vermaas, and Ibo van de Poel (Dordrecht: Springer Netherlands, 2015), 11–40, https://doi.org/10.1007/978-94-007-6970-0_3.

[128] Friedman, 'Value-Sensitive Design'.

One important challenge of VSD concerns the translation of abstract values into design requirements that can be used by designers. It has been proposed an approach where abstract values (e.g., privacy) are first translated into context-sensitive norms (e.g., do not collect personal data of kind Y in the domain X) and then into design requirements (the system should work without the need of personal data of the kind Y).[129]

Other challenges of VSD that are currently discussed and addressed in literature are the management of value tensions among different stakeholders and ethical approaches,[130] and the issue that values change over time.[131]

4.3. Value-Sensitive Design in Defence

Reviews of the state of the art, and future opportunities and limitations of the use of VSD in the military domain, with specific reference to AI-driven systems ('autonomous systems') have been conducted. [132] At least three specific challenges for the application of VSD in the military have been identified. The first challenge is finding effective methodologies to include critical stakeholders and engage them in the design process. This may be particularly difficult with military technologies due to the emotionally loaded topic, the lack of willingness of some key stakeholders (e.g., weapon developers to face the strong criticisms and fears of the public), and the secrecy of the military programmes. The second challenge is finding ways to elicit, identify, and aggregate public values about (future) military systems. One challenge here is that of combining 'top-down' approaches, more focused on values and principles solidified in existing institutions and endorsed by 'experts' such as lawyers or professionals, and bottom-up approaches, more focused on values and principles present in society broadly considered and visible in more informal spaces like public debates on social media platforms. The third challenge is connecting the values and principles emerging from a VSD process with existing systems design methodologies that system designers may be familiar with, to grant a smoother integration of these values and principles in the technical design process.

4.4. Design for Privacy, Privacy by Design (PbD)

A prominent example of the application of VSD in the civilian context is its implementation to safeguard privacy in Europe. VSD is a legal obligation under the EU GDPR as Article 25 of the GDPR states that data controllers are required to implement appropriate technical and organisational measures, such as pseudonymisation, which are designed to implement data protection principles, such as data minimisation, in an effective manner and integrate the necessary safeguards into the

[129] Ibo Van de Poel, 'Translating Values into Design Requirements', *Philosophy of Engineering and Technology* 15 (2013): 253–66, https://doi.org/10.1007/978-94-007-7762-0_20/FIGURES/3.
[130] Noëmi Manders-Huits, 'What Values in Design? The Challenge of Incorporating Moral Values into Design', *Science and Engineering Ethics* 17, no. 2 (12 June 2011): 271–87, https://doi.org/10.1007/s11948-010-9198-2.
[131] Ibo Van de Poel, 'Design for Value Change', *Ethics and Information Technology 2018 23:1* 23, no. 1 (26 June 2018): 27–31, https://doi.org/10.1007/S10676-018-9461-9.
[132] Boshuijzen-van Burken, C. 'Value Sensitive Design for autonomous weapon systems – a primer', *Ethics Inf Technol* 2023 **25**:11. https://doi.org/10.1007/s10676-023-09687-w

data processing. This approach is called data protection by design (DPbD) or simply privacy by design (PbD).[133]

It imposes obligations to consider different ways of collecting and processing personal data and subsequently chooses approaches that least interfere with the privacy of individuals. For instance, for business intelligence purposes, data do not always have to be available at an individual level. Processing data on aggregated levels or in anonymised ways may achieve the same results. These approaches are more privacy-friendly as employees of data controllers who are processing the data will see less privacy-sensitive information and in case of data breaches (e.g., hacks or leaks), it is less probable that privacy-sensitive information will be disclosed. Implementation of anonymisation and pseudonymisation tools, need-to-know access, role-based access controls, inference controls and audit systems are all examples of measures that can be designed into systems for collecting and processing personal data.

In line with privacy by design, the GDPR also imposes the obligation to consider privacy by default in Article 25 GDPR. Privacy by default aims to set defaults in technology in a privacy-friendly mode, for instance, opt-in instead of opt-out. Typically, social media will set any default settings to 'open for everyone', to increase the online visibility of their users and generate further activity on their platforms. However, the privacy by default requirement suggests that the default settings should be 'closed to everyone' unless a user indicates otherwise. Since only a few people change default settings, privacy by default can be an important tool to preserve and protect privacy.[134]

Although the principles of privacy by design and by default sound interesting, in practice they do not seem to be applied that often. These principles should be implemented before a new project starts, but in those stages, it may not seem to be the highest priority. Also, it can be complicated to develop strategies for privacy by design, as it requires sophisticated knowledge of the data processing and available approaches to render them more privacy-friendly. Even if experts are working on this, trade-offs between privacy and business interests may favour the latter rather than the former. As a result, only limited technological tools exist for implementing privacy by design and privacy by default. Nonetheless, tools continue to be developed under the moniker of Privacy-Enhancing Technologies[135].

[133] Cavoukian, A. (2010) Privacy by design: the definitive workshop. A foreword by Ann Cavoukian, *Identity in the Information Society*, 3(2), 247-251. For examples, see: Custers, B.H.M., Calders, T., Schermer, B., and Zarsky, T. (eds.) (2013) Discri*mination and Privacy in the Information Society: Data Mining and Profiling in Large Databases*, Heidelberg: Springer.

[134] Kesan, J.P., and Shah, R.C. (2006) Setting Software Defaults: Perspectives from Law, Computer Science and Behavioral Economics. *Notre Dame Law Review*, 583–634. Available online at http://law.bepress.com/uiuclwps/art54/. [53 p.]

[135] The Royal Society. (2019). 'Protecting privacy in practice: The current use, development and limits of privacy enhancing technologies in Data Analysis'. https://royalsociety.org/-/media/policy/projects/privacy-enhancing-technologies/protecting-privacy-in-practice.pdf

4.5. Design for Human Agency and Responsibility

4.5.1. Meaningful Human Control (MHC) and Human Oversight

Beyond privacy, important values that should be addressed by VSD in defence are those of human agency and responsibility. In defence, a lot of emphasis has recently been placed on 'meaningful human control' (MHC), the principle that humans not computers and AI should ultimately remain in control of, and thus morally responsible for, relevant decisions about (lethal) military operations.[136]

Following the legal and political debate on the regulation and governance of autonomous weapon systems (AWS), which are outside the scope of the ELSA Lab Defence – and the request of keeping them under meaningful human control, theoretical frameworks have been developed in design for meaningful human control.[137] The MHC framework has been further developed beyond the military.[138] Other converging proposals have tried to identify the relevant properties of meaningful human control to allow for its operationalisation into existing systems.[139] A group of designers and engineers have tried to further operationalised these properties in terms that are closer to engineering and design practices.[140] Meaningful human control has also been presented as a key element of a strategy to prevent various kinds of responsibility gaps with AI (see section 3.3 above).[141]

A framework to design for comprehensive human oversight has also been proposed that further develops the above-mentioned idea of MHC as crucially depending not on direct human control from an operator but on the presence of a complex structure of oversight over and accountability for the system operation. [142] Such an oversight structure would operate at various levels: the technical level (system design), the socio-technical level (human-machine interaction), and the governance/institutional level (e.g., operation review, accountability system). It should be applied across time, i.e., before, during and after deployment.

---

[136] Article 36. (2015). Killing by Machine: Key Issues for Understanding Meaningful Human Control. Available at: https://article36.org/wp-content/uploads/2020/12/KILLING_BY_MACHINE_6.4.15.pdf

[137] Santoni de Sio and van den Hoven, 'Meaningful Human Control over Autonomous Systems: A Philosophical Account'.

[138] Mecacci et al. 2024.

[139] Heather M. Roff and Richard Moyes, 'Meaningful Human Control, Artificial Intelligence and Autonomous Weapons', 2016; Daniele Amoroso and Guglielmo Tamburrini, 'What Makes Human Control Over Weapon Systems "Meaningful"?', *ICRAC Working Paper Series #4*, 2019.

[140] Luciano Cavalcante Siebert et al., 'Meaningful Human Control: Actionable Properties for AI System Development', *AI and Ethics* 3, no. 1 (1 February 2023): 241–55, https://doi.org/10.1007/s43681-022-00167-3.

[141] Santoni de Sio and Mecacci, 'Four Responsibility Gaps with Artificial Intelligence: Why They Matter and How to Address Them'.

[142] Ilse Verdiesen, Filippo Santoni de Sio, and Virginia Dignum, 'Accountability and Control Over Autonomous Weapon Systems: A Framework for Comprehensive Human Oversight', *Minds and Machines* 31, no. 1 (1 March 2021): 137–63, https://doi.org/10.1007/S11023-020-09532-9/FIGURES/6.

4.5.2. Explainable AI (XAI)

There has been a long-term debate on how AI can be meaningfully controlled by humans and who is responsible for AI technology in case things go wrong. One of the key elements in controlling AI is that of transparency, as it can be argued that AI can be hard to control once people no longer understand what it does. The same applies to allocating responsibilities, which can only be done properly when it is clear how the AI technology works. AI should not work as a 'black box'.[143] In the last decade, the focus of this debate has shifted from transparency towards explainability.[144] Instead of exactly seeing what AI does (i.e., transparency), it is sufficient to understand what AI does (i.e., explainability). Explainability can be applied in retrospect, such as reverse engineering how AI came to a decision and does not require a priori transparency, which may be hard for human intuition anyway. Although for many types and applications of AI explainability may be unnecessary, for critical applications, such as defence, it is essential for users to understand, trust, and manage the AI systems they deploy.[145]

The focus on developing explainable AI is intended as a step to move from methods incorporating values and principles into the design of AI to building mechanisms that demonstrate responsible behaviour (responsible AI). These mechanisms are not merely technological, they can be found in three components of AI systems.

Firstly, institutional mechanisms intend to shape or clarify the incentives of people involved in AI development, including their efforts to ensure safe, secure, fair and privacy-preserving AI systems. Mechanisms include third-party auditing, red teaming exercises, bias and safety bounties, and sharing of AI incidents.

Secondly, software mechanisms intend to increase understanding and oversight of the behaviour and characteristics of AI systems. Mechanisms include audit trails, interpretability and privacy-preserving machine learning.

Thirdly, hardware mechanisms can play a strong role in substantiating claims about privacy and security and enable transparency. Mechanisms for this include secure hardware for machine learning, high-precision compute measurement, and compute support for academia.

Responsible AI is an often-used term, but something of an empty shell, sometimes used for window-dressing.[146] The XAI approach addresses this by concrete measures to increase understandability.

---

[143] Pasquale, F. (2015). *The black box society*. Harvard University Press.

[144] Brundage, M., Avin, S., Wang, J., Belfield, H., Krueger, G., Hadfield, G., ... & Anderljung, M. (2020) Toward trustworthy AI development: mechanisms for supporting verifiable claims. <https://arxiv.org/pdf/2004.07213.pdf &hellip>.

[145] Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G. Z. (2019). XAI—Explainable artificial intelligence. *Science robotics*, *4*(37).

[146] Boulanin, V., Lewis, D.A. Responsible reliance concerning development and use of AI in the military domain. *Ethics Inf Technol* 25, 8 (2023).

This, in turn, can increase reliance, trust, and control concerning these systems. Similar to the VSD approach,[147] the XAI approach focuses on implementing these measures in the technology design.

There are several ways in which AI models can be explained.[148] One way is by simplifying an AI system via approximation.[149] A second way is by explaining the features of an AI system.[150] A third way is by providing visualisations that are easier to understand for humans.[151] A fourth way is to provide local explanations, with a focus on explaining specific input and output relations without a need to explain the entire complexity of an AI model.[152]

4.5.3. Contestability by Design

Taking its cue partly from the right to contest decisions based solely on the automated processing of data pursuant to Article 22(3) GDPR, contestability-by design is an approach to designing autonomous systems in such a way as to ensure that their algorithmic outcomes can be challenged by a directly or indirectly affected individual. Contestability "helps to protect against fallible, unaccountable, illegitimate, and unjust automated decision-making, by ensuring the possibility of human intervention as part of a procedural relationship between decision subjects and human controllers".[153] Having not only the right, but also the technical-practical ability to contest predictive results of AI processes is critical to prevent harm.

To ensure contestability as the outcome of the design process of an AI system and not simply as a reactive right, it has been proposed that a framework consisting of "five system features and six development practices" be implemented.[154] The five system features are built-in safeguards against harmful behaviour; interactive control over automated decisions; explanations of system behaviour; human review and intervention requests; and tools for scrutiny by subjects or third parties. The six development practices are: ex-ante safeguards; agonistic approaches to machine learning

[147] See, Section 4.3.

[148] Angelov, P. P., Soares, E. A., Jiang, R., Arnold, N. I., & Atkinson, P. M. (2021). Explainable artificial intelligence: an analytical review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *11*(5).

[149] Tritscher, J., Ring, M., Schlr, D., Hettinger, L., & Hotho, A. (2020). Evaluation of post-hoc XAI approaches through synthetic tabular data. In International symposium on methodologies for intelligent systems (pp. 422–430). Springer.

[150] Chen, H., Lundberg, S., & Lee, S.-I. (2019). Explaining models by propagating Shapley values of local components. arXiv preprint arXiv: 1911.11888; Pedreschi, D., Giannotti, F., Guidotti, R., Monreale, A., Ruggieri, S., & Turini, F. (2019). Meaningful explanations of black box AI decision systems. In Proceedings of the AAAI conference on artificial intelligence, Vol. 33, p. 9780–9784.

[151] Chattopadhay, A., Sarkar, A., Howlader, P., & Balasubramanian, V. N. (2018). Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks. In 2018 IEEE Winter conference on applications of computer vision (WACV) p. 839–847.

[152] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE international conference on computer vision, p. 618–626.

[153] Kars Alfrink and others, 'Contestable AI by Design: Towards a Framework' [2022] Minds and Machines <https://doi.org/10.1007/s11023-022-09611-z> accessed 15 June 2023.

[154] Alfrink and others (n 92).

development; quality assurance during development; quality assurance after deployment; risk mitigation strategies; and third-party oversight.

Contestation should not only come into play *ex post*, i.e., after the system has already produced potentially injuring output. Instead, contestability *by design* means that the ability to find a remedy against the system in a meaningful way can only be achieved if effective points for human intervention are deliberated and implemented during the development process.[155] One of the critical features of such contestability-by-design is the default capability of affected individuals to grapple with the substance of the AI system's decision, which is a precondition for the explainability of algorithmic processes.[156] Contestation is only truly possible if a natural person can understand the inner workings of the system and the substantial reasons for certain harmful outcomes. For this to happen, in turn, the subject of a decision made by an autonomous agent must have been aware that automated processing occurred.[157] Thus, the approach is to be understood as a framework to design AI systems that enable human intervention and actual challenging of decisions by involving human agency as early in the development process as possible, ideally through iterative stakeholder participation.[158]

4.6. Other Engineering Approaches to Design for Human Agency

4.6.1. Cognitive and Socio-Cognitive Engineering

Next to VSD, there are other approaches for designing for human agency in complex socio-technical systems. The field of cognitive engineering emerged in the early 1980s as a response to the observation that emergent complex systems built around human-computer interaction require novel approaches to ensure the controllability of safety-critical systems such as nuclear power plants or commercial aircraft.[159] Originally, cognitive engineering mainly aimed at improving the alignment between the human operator and the system for the sake of work efficiency, acknowledging that increasingly powerful technology renders humans the most likely limiting factor.[160] Conceptually, it can be described as a type of applied cognitive science, in the sense that it builds on the discipline's

[155] Marco Almada, 'Human Intervention in Automated Decision-Making: Toward the Construction of Contestable Systems', *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law* (Association for Computing Machinery 2019) <https://dl.acm.org/doi/10.1145/3322640.3326699> accessed 15 June 2023.
[156] Claudio Sarra, 'Put Dialectics into the Machine: Protection against Automatic-Decision-Making through a Deeper Understanding of Contestability by Design' (2020) 20 Global Jurist <https://www.degruyter.com/document/doi/10.1515/gj-2020-0003/html?lang=en> accessed 15 June 2023.
[157] Almada (n 93).
[158] See next section, "Participatory Design and Evaluation Methods".
[159] John R Gersh, Jennifer A McKneely and Roger W Remington, 'Cognitive Engineering: Understanding Human Interaction with Complex Systems' (2005) 26 Johns Hopkins APL Technical Digest 377.
[160] ibid.

findings to design and construct machines.[161] It understands the engineering process as a series of trade-offs that inevitably have to be made when the human operator's psychological variables and the machine's technical-physical variables are sought to be aligned in a way that allows for efficient yet safe operating, taking into account any feedback loops involved.[162] In this sense, cognitive engineering differs substantially from the aforementioned design approaches in that it does not in itself encode ethical, political, or other external preferences, but only aims at ensuring that whatever such preferences might be, a human user must be able to translate them into system outcomes. In other words, it is about the 'effective governance' of machines.[163]

With the rise of AI, research in cognitive engineering has gradually begun to expand its scope and focus on questions of human supervision of autonomous systems.[164] Specifically, it has been argued that "the foundation for a precise, comprehensive and robust definition of [meaningful human control] is found in the discipline of cognitive engineering, whose primary focus is on the interaction between automation and humans, particularly in complex and dynamic domains".[165] The concept of 'function allocation', as developed within cognitive engineering, provides an adequate theoretical framework for conceptualising the control of autonomous agents that are employed in public security scenarios. With reference to earlier work,[166] it has been put forward that five requirements for effective function allocation are: (1) each agent must be allocated functions that it is capable of performing, (2) each agent must be capable of performing its collective set of functions, (3) the function allocation must be realisable with reasonable teamwork, (4) the function allocation must support the dynamics of the work, and (5) the function allocation should be the result of deliberate design decisions.[167]

This list of requirements forms the basis for implementing policy choices and preferences. In a military context, this can be the rules of armed conflict, such as the principles of distinction and proportionality. Making use of the function allocation form when designing the autonomous system thus helps to guarantee that the interactive relationship between the human operator and machine is capable, on a technical level, of adhering to the previously agreed-upon ethical-legal framework.[168] In this sense, cognitive engineering should be understood as a design-based approach, providing the necessary technical grounding for the effective governance and regulation of autonomous agents.[169]

---

[161] Donald Arthur Norman, 'Cognitive Engineering' in Stephen W Draper and Donald Arthur Norman (eds), *User Centered System Design: New Perspectives on Human-Computer Interaction* (1986). 31
[162] ibid.
[163] Marc Canellas and others, 'Framing Human-Automation Regulation: A New Modus Operandi from Cognitive Engineering' (23 March 2017) <https://papers.ssrn.com/abstract=3567175> accessed 17 May 2023.
[164] Gersh, McKneely and Remington (n 53).
[165] Marc Canellas and Rachel Haga, 'Toward Meaningful Human Control of Autonomous Weapons Systems Through Function Allocation' (26 October 2015) <https://papers.ssrn.com/abstract=2927702> accessed 17 May 2023.
[166] Karen M Feigh and Amy R Pritchett, 'Requirements for Effective Function Allocation: A Critical Review' (2014) 8 Journal of Cognitive Engineering and Decision Making 23.
[167] Canellas and Haga (n 59).
[168] ibid.
[169] Canellas and others (n 57).

Socio-cognitive engineering (SCE) merges principles from different disciplines, such as the social sciences, cognitive psychology, and engineering to help understand human behaviour in social systems and then create human-centred technology.[170] This approach was developed after the emergence of cognitive engineering in the 1980s, adding the 'socio' element. SCE acknowledges that human behaviour is not exclusively driven by individual choices or preferences but is also influenced by social and cognitive factors. In other words, it recognises that people's actions are shaped by social norms, beliefs, perceptions, and the influence of others in our environment. By understanding these underlying processes, socio-cognitive engineering aims to engineer social systems that promote desired behaviours, facilitate positive outcomes, and address complex societal challenges.[171]

The distinctive element of SCE is, therefore, the integration of the introduction of social and sociological aspects to cognitive engineering in the design of technology. Understanding human behaviour in social systems, and combining knowledge from different disciplines, it allows for the creation of human-centric technology.

4.6.2. Human-Machine Teaming

One way to interpret and operationalise the requirements of meaningful human control is that of creating a conceptual and design framework "to build and then utilise a partnership between people and AI so that each party utilises its strengths to achieve a shared goal".[172] The use of the word 'team' thereby implies the purposeful interaction between agents, which means that on the part of the machine, there needs to be some degree of autonomy realised for the notion to be meaningful.[173] If that is the case, the individual components of the team perform complementary, non-redundant tasks to achieve a shared, previously agreed goal within an organisational structure and situational constraints that together prevent or limit unpredictable results despite the individual members' autonomous behaviour. The teammates share situational awareness and are capable of learning from each other and adapting to each other's reactions and actions.[174] Building on this set of preconditions, the concept of MHC can then be utilised as a toolkit to 'bridge the gap' between an organisation's ethical or legal framework and its application in situational settings that require ongoing interaction between human agents and autonomous systems, thereby ensuring that the former remain in control.[175]

---

[170] Neerincx, M. A., and Lindenberg, J. (2008). "Situated cognitive engineering for complex task environments," in Naturalistic Decision Making and Macrocognition, eds J. M. C. Schraagen, L. Militello, T. Ormerod, and R. Lipshitz (Ashgate Publishing, Ltd.), 373.

[171] https://www.frontiersin.org/articles/10.3389/frobt.2019.00118/full#B55.

[172] Aiden Warren and Alek Hillas, 'Friend or Frenemy? The Role of Trust in Human-Machine Teaming and Lethal Autonomous Weapons Systems' (2020) 31 Small Wars & Insurgencies 822.

[173] Brill and others (n 78).

[174] ibid.

[175] Carol J Smith, 'Designing Trustworthy AI: A Human-Machine Teaming Framework to Guide Development' (arXiv, 8 October 2019) <http://arxiv.org/abs/1910.03515> accessed 14 June 2023.

Human-machine teaming provides a conceptualisation for a socio-technical design approach that allows for the implementation of ethical-legal frameworks in AI-supported technologies such as the notion of 'meaningful human control' of autonomous agents in military contexts.[176] Rejecting frequent suggestions that meaningful human control can be realised by simple adherence to the 'human-in-the-loop/human-on-the-loop' model, it has been contended that any autonomous agent must be conceived and designed from the start as a 'teammate' of human operators to achieve truly collaborative arrangements. Supporting humans and AI systems as teammates requires "a carefully designed system. Critical team functions include sharing situation awareness, understanding each other's role, managing interdependencies, aligning goals and plans".[177]

Different possible configurations of collaborative human-machine interaction can then be outlined, for example, by means of 'team design patterns' that schematically depict the allocation of critical and non-critical functions between humans and autonomous agents. The human-machine teaming design method is not to be understood as a sufficient and all-encompassing approach to solving issues surrounding meaningful human control of autonomous agents by itself but is rather to be employed in combination with other pertinent methods to 'iteratively design for MHC'.[178]

The MHC approach is by definition highly contextual and, thus, not amenable to standardised, one-size-fits-all solutions. Instead, the specific requirements for human-machine interaction in a concrete use case are determined by interdependent factors pertaining to the involved human operator(s), the autonomous system itself, and the respective task environment.[179] To enable an already constructed autonomous system to meaningfully interact with human operators as part of a team, it has been proposed that a modular method to integrate humans and machines via a dedicated layer for social interaction (SAIL = Social AI Layer) that utilises a specific linguistic framework (HATCL = Human-Agent-Teaming Communication Language).[180] In doing so, the SAIL serves to illustrate what a particularly detailed and sophisticated application of the general idea of human-machine teaming might look like in practice.

Some countries have already endorsed the human-machine teaming framework in their thinking on the use of AI in the military domain. In 2018, the UK Ministry of Defence issued an extensive report on this topic.[181] Among other aspects, it focuses on the critical question of how human operators can put sufficient trust in the autonomous agents they are supposed to form teams within the near future. The report focuses on four 'fundamental factors' that determine trust in this regard: mechanical

---

[176] Jurriaan van Diggelen et al., 'Designing for Meaningful Human Control in Military Human-Machine Teams', in *Research Handbook on Meaningful Human Control of Artificial Intelligence Systems*, ed. Giulio Mecacci et al. (Cheltenham, UK: Edward Elgar Publishing, 2024).
[177] ibid.
[178] ibid.
[179] Bob van der Vecht and others, 'SAIL: A Social Artificial Intelligence Layer for Human-Machine Teaming' in Yves Demazeau and others (eds), *Advances in Practical Applications of Agents, Multi-Agent Systems, and Complexity: The PAAMS Collection* (Springer International Publishing 2018).
[180] ibid.
[181] UK Ministry of Defence, 'Joint Concept Note 1/18: Human-Machine Teaming' (2018).

understanding, predictability, familiarity, and context.[182] In the Netherlands, also a variation of the MHC concepts is used in a 2019 statement to the UN Group of Governmental Experts on Lethal Autonomous Weapons Systems, emphasising the need for meaningful human control by considering that effective human-machine teaming may allow for the optimal utilisation of technological benefits, such as precision, speed and reliability without sacrificing the robustness and flexibility of human intelligence.[183]

4.6.3. Team Design Pattern Engineering

In team design pattern engineering, the focus is not only on the technical design, but also on the human and social dynamics within a team. Within the team, this includes human-machine interaction. In this sense, the approach is to be understood as a derivation from the more general framework of MHC.

Team design pattern engineering is the identification and selection of appropriate design patterns that align with the team's goals, project requirements, and development context. This involves analysing the team's needs, understanding the problem domain, and considering factors such as scalability, maintainability and extensibility. Once the relevant design patterns are identified, the team can employ them as a shared language and framework for collaboration. Design patterns provide a common vocabulary and a set of well-defined solutions, facilitating communication among team members and promoting a shared understanding of the software architecture.[184]

By incorporating team design pattern engineering, the idea is that teams can enhance their productivity, code quality, and overall performance. It fosters a collaborative and structured approach to design and encourages the sharing of best practices and lessons learned among team members. Ultimately, team design pattern engineering is supposed to promote a more efficient and cohesive team environment, leading to successful software development outcomes.

4.7. Engaging Society in the Design of Military Systems

Value-Sensitive Design and other methodologies presented so far require the involvement of stakeholders in various degrees. The following subsections offer a critical presentation of some recent attempts to better realize this involvement.

---

[182] Ibid.

[183] Statement of the Netherlands Delivered at the Group of Governmental Experts on LAWS, Geneva, 26 April 2019, p. 3.

[184] Jurriaan van Diggelen and Matthew Johnson, 'Team Design Patterns', *Proceedings of the 7th International Conference on Human-Agent Interaction* (Association for Computing Machinery 2019) <https://dl.acm.org/doi/10.1145/3349537.3351892> accessed 23 June 2023.

4.7.1. Inclusion of Workers in Design: Scandinavian Participatory Design

One of the first systematic attempts to realise the ideal of the direct involvement of stakeholders in technological design was called participatory design. The name originally covered a series of experimental projects in Norway and Sweden from the 1970s.[185] For this reason, this approach is now sometimes referred to as Scandinavian participatory design, to distinguish it from the other, more recent, approaches developed and discussed below in this section ('new participatory design'). In Scandinavian participatory design, workers were invited to meetings where they would discuss with managers and engineers the introduction of new technologies in the workplace – typically computers and information systems. The idea was that workers would bring their experience and values to the table and, therefore, influence the development and introduction of the technologies they would eventually be asked to use. In their intentions, this aligns with the ideals of both Responsible Innovation[186] and VSD. Participatory design in the workplace has been praised by some as an important step towards democratising technology.[187]

Participatory design also faced some important issues. Most importantly, it was difficult to ensure that workers had the real power to make their voices and ideas matter in the final managerial choices and even more in the design process.[188] Similar criticisms have been raised about stakeholder involvement in VSD.[189] The risk of stakeholders not being able to freely contribute is particularly high in hierarchical contexts like workplaces, including the defence domain, where the workforce may feel the pressure to not challenge the management's decisions and plans. Also, well-intentioned managers and engineers may be unable to effectively communicate the technical aspects of their plans to workers and workers may, in turn, have issues communicating their fears and concerns.

4.7.2. Critical, Speculative, Social, (New) Participatory Design

In response to the limitations of VSD and Scandinavian participatory design, other design methodologies have since emerged that focus on addressing power dynamics in stakeholder involvement. These methodologies aim to address a form of power shaping technological development: the power of dominant cultures, visions and narratives. To ensure an open and plural technological process, the domination of certain stakeholders in the design process and certain dominant narratives that make those stakeholders' values accepted in society need to be addressed.

---

[185] Peter M. Asaro, 'Transforming Society by Transforming Technology: The Science and Politics of Participatory Design', *Accounting, Management and Information Technologies* 10, no. 4 (October 2000): 257–90, https://doi.org/10.1016/S0959-8022(00)00004-7.

[186] Richard Owen, René von Schomberg, and Phil Macnaghten, 'An Unfinished Journey? Reflections on a Decade of Responsible Research and Innovation', *Journal of Responsible Innovation* 8, no. 2 (4 May 2021): 217–33, https://doi.org/10.1080/23299460.2021.1948789.

[187] Asaro, 'Transforming Society by Transforming Technology'.

[188] Asaro.

[189] Alan Borning and Michael Muller, 'Next Steps for Value Sensitive Design', in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY, USA: ACM, 2012), 1125–34, https://doi.org/10.1145/2207676.2208560.

'Critical design' is a term coined in the mid-1990s to describe a design practice that "uses products to ask questions and raise issues in society and culture."[190] It typically uses prototype products to prompt questions about possible and (un)desirable futures that technological development can lead to. In a way, it is not far from science fiction and other forms of art reflecting on the possible impact of technology on humanity and society. It asks not only what kind of technological products stakeholders want or need, but more broadly, what kind of society do people and citizens want to live in, and what the role of technology should be. In critical design, products are designed with the primary intention to push the limits of the lived experience of those interacting with the artefact. The relation between stakeholder participation and artefacts is somehow reversed. Here, the designed object or system is meant to work for eliciting the stakeholder's understanding and political engagement, rather than the stakeholder's engagement being functional to improving the product. Rather than just articulating ethical values or designing ethics into technology, critical design aims at prompting a societal and political debate around the desirability of certain existing trends in the development of technology.

In partial overlap with critical design, but possibly with more open-ended and explorative attitude, is the bordering field of 'speculative design', where artefacts, settings, and experiences are designed to prompt people's imagination about alternative presents and speculative futures.[191]

Recently, some scholars tried to connect the traditions of social and participatory design with that of critical and speculative design.[192] These scholars put forward that these two approaches can fruitfully complement each other. Critical and speculative design have the power to spark reflection and imagination and open new visions, but they have been traditionally practised by professional designers in relative isolation from other social groups and communities and with a relatively low societal impact. Social and participatory design have more directly addressed communities and their present issues but possibly without the political, philosophical, and creative breadth of critical design. Combining the two may offer ways to create original design experiments that both prompt a reflection on new possible desirable (technological) futures and empower local communities in addressing their issues starting from their visions and expertise. The need for a real empowering of communities in design practices is also stressed in the new movement of design justice,[193] that has criticised previous attempts at involving stakeholders such as those of VSD and Scandinavian participatory design as a form of qualitative data extraction that largely benefit professional designers and policymakers rather than the social communities that will be affected by technology.

---

[190] Anthony Dunne and Fiona Raby, *Speculative Everything: Design, Fiction, and Social Dreaming* (Cambridge, Massachusetts London: MIT Press, 2013).

[191] James Auger, 'Speculative Design: Crafting the Speculation', *Digital Creativity* 24, no. 1 (1 March 2013): 11–35, https://doi.org/10.1080/14626268.2013.767276.

[192] Carl DiSalvo, *Design as Democratic Inquiry: Putting Experimental Civics into Practice* (Cambridge, Massachusetts: The MIT Press, 2022); Laura Forlano, 'Decentering the Human in the Design of Collaborative Cities', *Design Issues* 32, no. 3 (1 July 2016): 42–54, https://doi.org/10.1162/DESI_a_00398.

[193] Sasha Costanza-Chock, *Design Justice: Community-Led Practices to Build the Worlds We Need*, Information Policy (Cambridge, Massachusetts: The MIT Press, 2020).

4.7.3. Enhancing Societal Reflection on Military Technologies: Critical and Speculative Design

The methodology of VSD arguably reflects the political ideal of democratic deliberation.[194] According to deliberative democracy theory, the involvement of citizens in public decision-making increases the quality and legitimacy of the decisions (in addition to respecting people's equal right to participate). Similarly, the systematic involvement of so-called stakeholders in the design process increases the quality of the technological product and legitimacy of the innovation process. Indeed, as emphasised also in the literature on responsible innovation,[195] the involvement of stakeholders may allow for better responsiveness of technological development to societal values and interests,[196] and may help anticipating and preventing unwanted outcomes thanks to the knowledge and expertise offered by stakeholders.[197]

The Scandinavian participatory design approach, in which workers contribute to the design of the technical systems they use, was seen as a way to ensure that these tools would 'work better' by being more aligned to the users' will and needs.[198] This approach has been developed for other contexts, beyond the workplace, under the label of user-centred or human-centred design.[199]

The concept of 'stakeholder involvement' has recently been criticised. Design scholars have noticed that 'stakes' and 'stakeholders' were conceptualised in the realm of organisations whose activities are classified as 'projects'.[200] However, as the focus of codesign research is shifting to broader societal issues and public concern, the rhetoric and assumptions about stakes and stakeholders are being challenged. Moving toward a more experimental approach to democracy, it has been claimed that this kind of assumed a priori juxtaposition of groups defined by a more or less singular 'stake' may hinder the formation of new issues and so-called emerging 'publics', i.e., groups of people collectively engaging on a certain social or political issue.[201] It is proposed by these scholars to think of the design process not as much as a deliberation and negotiation between already formed groups, interests and opinions, but rather as the space where these groups, interests and opinions are formed.[202] This is "a fundamentally experimental attitude to the epistemology of the collaborative encounter."[203] In this respect it is interesting to note that a study showed that, perhaps contrary to what may be expected, military experts moral intuitions and concerns about autonomous weapon

---

[194] Santoni de Sio, *Human Freedom in the Age of AI*, 201–30.
[195] Owen, von Schomberg, and Macnaghten, 'An Unfinished Journey?'
[196] Jack Stilgoe, Richard Owen, and Phil Macnaghten, 'Developing a Framework for Responsible Innovation', *Research Policy* 42, no. 9 (2013): 1568–80, https://doi.org/10.1016/j.respol.2013.05.008.
[197] Jeroen van den Hoven, 'Value Sensitive Design and Responsible Innovation', in *Responsible Innovation: Managing the Responsible Emergence of Science and Innovation in Society*, ed. Richard Owen, John Bessant, and Maggy Heintz (John Wiley & Sons, Ltd, 2013), 75–83, http://onlinelibrary.wiley.com/doi/10.1002/9781118551424.ch4/summary.
[198] Asaro, 'Transforming Society by Transforming Technology'.
[199] Guy A. Boy, *The Handbook of Human-Machine Interaction: A Human-Centered Design Approach* (Boca Raton, Florida: CRC Press, 2017).
[200] Thomas Binder et al., 'Democratic Design Experiments: Between Parliament and Laboratory', *CoDesign* 11, no. 3–4 (2 October 2015): 152–65, https://doi.org/10.1080/15710882.2015.1081248.
[201] Binder et al.
[202] Binder et al., 'Democratic Design Experiments', 162.
[203] Binder et al., 161.

systems do not differ much from those of laypeople.[204] Therefore, a design approach that would assume an interaction between military people as stakeholders in favour and laypeople as stakeholders against weapon systems with autonomous capabilities would prevent the formation of more diverse and articulated groups and opinions.

Following the idea of 'critical design',[205] spaces and activities can be designed where expert and non-experts can be involved in experiences and activities – not just talking, debate and deliberating – through which they may articulate their visions, fears and expectations towards future technologies in the military. This may allow people to develop a more articulated position on such issues, reshuffling intuitions and biases, problematising dominant narratives – positive or negative - about military technologies. This may also allow people who do not want to contribute to (certain kinds of) technological developments – nor to the deliberation about their design and use – to have their voiced heard, amplified, potentially better understood and endorsed by others.

4.7.4. Giving More Power to People in Military Design: Social and (New) Participatory Design

Both VSD and critical design have been criticised for not giving stakeholders and people sufficient influence on the design process according to their needs, preferences and visions. As for VSD, it has been observed that notwithstanding their best intentions, designers still tend to play a dominant role in the process. They often influence the identification of relevant values by stakeholders, for instance, by working with a predefined list of values, which may make them ignore other values held by stakeholders.[206] Also, researchers or designers always maintain the power to present the results of their empirical investigation in their own words. This means that stakeholder contributions are often interpreted and paraphrased by the researchers, with the risk of researchers projecting their own views and perspectives on this.[207] There is an ongoing debate as to whether these are intrinsic limitations of the value-sensitive design, or they can be addressed by revising its methodology.[208]

As for critical design, it is clear that imagination, one-off design prototypes, or artistic projects conveying radical critique are not enough. Imagination and critique should be somehow embedded and reflected in the design process to be effective and produce lasting consequences. To this goal, design theorists have proposed to combine critical design with the tradition of social and (new) participatory design.[209]

---

[204] Ilse Verdiesen, 'Agency Perception and Moral Values Related to Autonomous Weapons: An Empirical Study Using the Value-Sensitive Design Approach' (2017), https://repository.tudelft.nl/islandora/object/uuid:7cc28c2e-69d9-45f3-9c87-51e8281c32b0?collection=education.

[205] See, Subsection 4.7.2.

[206] Christopher A. Le Dantec, 'Situated Design: Toward an Understanding of Design through Social Creation and Cultural Cognition', in *Proceedings of the Seventh ACM Conference on Creativity and Cognition* (C&C '09: Creativity and Cognition 2009, Berkeley California USA: ACM, 2009), 69–78, https://doi.org/10.1145/1640233.1640247.

[207] Borning and Muller, 'Next Steps for Value Sensitive Design'.

[208] Borning and Muller.

[209] Forlano, 'Decentering the Human in the Design of Collaborative Cities'; DiSalvo, *Design as Democratic Inquiry: Putting Experimental Civics into Practice*.

Warnings on the risk of stakeholder participation all-too-easily becoming 'extractive' rather than genuinely participative have been made.[210] Such warnings emphasise the importance of collaborating with communities to recognise and value their expertise and to expand this existing expertise in the communities' interest, rather than seeing stakeholders as a source of knowledge and information to be used in the interests of or companies, governments, and others.

The idea of 'design justice' captures the spirit of social and (new) participatory design. Here communities take centre-stage in the design process, and the designer's role shifts. Designers are no longer technical experts trying to combine in their projects the knowledge, interest, and expertise of different stakeholders. They are more like facilitators of a collaboration between societal actors already involved in the creation of their own products and processes.

In addition to supporting a broader, more productive, and more creative democratic participation in the deliberation about future technology, these approaches also aim to promote social justice and distribution of power, by giving communities and people the power to develop skills and own projects they deem valuable, as opposed to be involved as 'participants' in projects devised and run by other, often more powerful actors, such as governments or companies.

One may wonder to what extent such an approach may be applied in the military domain, considering the high level of technological complexity and sensitivity of military technology and operations. Without underestimating the magnitude and complexity of the challenge, some preliminary considerations in favour of this approach may be proposed.

First, not all military technologies are so complex that they require big technological and research infrastructure for their production, as demonstrated by the proliferation of the production of systems for military use by non-state groups. And even when such infrastructures are required, providing such support would precisely be the role of professional designers and engineers involved. Social and (new) participatory design do not require people and communities to design on their own infrastructure, but rather to be the at the centre of the process.

Relatedly, the distinction between military and non-military technology is blurring as demonstrated by the case-study of cognitive warfare, in which the technology used for strategic and military purposes is non-military in character, e.g., social media platforms and AI-powered bots. Therefore, whether they realise it or not, people and communities of 'non-experts' are *de facto* already involved in the use and development of sensitive technologies and systems, so they should rather be involved more explicitly in the design, development and use of these systems.

Finally, it is often claimed that the most difficult and critical task for military organisations is 'winning hearts and minds' of the people they want to serve. The cognitive warfare case study shows that this is more and more the case. Looking at social and (new) participatory design approaches, it can be suggested that giving people more power and recognition in the development of defence-sensitive systems – whether military in nature or not – may be a more just but also a more effective way to have them engaged in the promotion of the security of their country.

[210] Costanza-Chock, *Design Justice*.

4.8. Design for Emerging Human-Technology Interactions

A recurring criticism of VSD is that it sees the human element as something distinct from technology and design, something to which design should adapt. Scholars in sociology of science[211] and post-phenomenological approaches in the philosophy of technology[212] have convincingly argued that this picture is inaccurate and misleading. Human values and technology tend to shape each other dynamically and continuously.

Rather than taking human values and needs for granted and designing products around them, the design process should create spaces, objects and systems that make people and communities gather, discuss and develop awareness of their relationship with technology, and produce new visions and values for possible future interactions with technology. The above-mentioned design approaches of critical, speculative, social, (new) participatory design aim to facilitate the emergence of *new* values and visions through the systematic design of interactions between professional designers and other people and groups and between these and new emerging systems and technologies.

In the scientific and technological communities, the most systematic attempts to design for these emerging interactions have taken place in the field of human-computer-interaction (HCI), human-machine interaction (HMI) and human-robot interactions (HRI). An open challenge is connecting these existing technological practices with emerging forms of more explicitly ethically and societally driven forms of design, which are discussed in this section.

4.8.1. Coactive Design

Early robotics and software agent researchers were heavily influenced by scenarios in which autonomous technologies were thought to 'replace' human interaction, limiting the need to take into account the 'social' aspects of working together.[213] Designing human-robot systems using traditional methods typically involves task allocation and decomposition. The best-known use of this strategy is supervisory control,[214] in which tasks are delegated to one or more machines, and their performance is subsequently observed. One of the problems with this method is that a person or machine's appropriateness for a given task may change over time and in various contexts.

Coactive design is a method for addressing the various roles that humans and robots play as the use of robots spreads into new, intricate fields. To describe a method for designing human-robot interaction (HRI) that uses interdependence as the main organising principle for people and robots cooperating in joint action, the term 'coactive design' was developed.[215]

---

[211] Bruno Latour, *We Have Never Been Modern*, 3. print. (Cambridge, Mass: Harvard Univ. Press, 1994).
[212] Peter-Paul Verbeek, *What Things Do: Philosophical Reflections on Technology, Agency, and Design* (University Park, Pennsylvania: Penn State University Press, 2005).
[213] https://dl.acm.org/doi/pdf/10.5898/JHRI.3.1.Johnson.
[214] Sheridan, T. B. (1992). Telerobotics, automation, and human supervisory control. Cambridge, MA: MIT Press.
[215] Johnson, M., Bradshaw, J., Feltovich, P., Jonker, C., van Riemsdijk, B., & Sierhuis, M. (2011). The fundamental principle of coactive design: Interdependence must shape autonomy. In M. De Vos, N. Fornara, J. Pitt, & G.

The term 'coactive' is intended to emphasise the reciprocal and mutually constraining nature of actions and outcomes that are conditioned by coordination, in addition to suggesting that two or more parties are involved in the activity. Systems where people and machines work together are the subject of coactive design. The word 'joint activity' refers to the type of activity, and 'coactive design' refers to the process of designing in a way to achieve successful joint activity. To create systems that support these relationships and help designers achieve the goals of coordination, cooperation, and teamwork, coactive design aims to assist designers in identifying interdependent relationships in a collaborative activity.

4.8.2. Adaptive and Adaptable Automation

Another engineering approach that aims to tackle dynamic contexts and enable effective human-machine collaboration is adaptive automation. In adaptive automation, work is allocated in real-time to a person or machine, depending on predefined factors such as the environment, the task, or operator states.[216] In this way, adaptive automation acknowledges that for some tasks, a person may be better suited, while other tasks are better handled by machines.[217] When a change in the level of automation (e.g., from full automation to manual control) is triggered by the machine or autonomously, this is called adaptive automation. When the change in automation is triggered by a person, this is called adaptable automation.

Although adaptive automation was originally designed with the goal of optimising system performance,[218] recent research shows that it can be applied to safeguard ELSA.[219] Through adaptive automation, it is possible to have machines carry out low-level tasks autonomously but (request to) switch to manual control when morally sensitive situations occur. This can be used in the military domain e.g., to have a drone perform autonomous area surveillance but ask for human input when unexpected things occur such as civilians being present in the area. This method of applying adaptive automation shows that it is possible to connect technical practices from HCI, HMI, and HRI to explicit ethical and social design approaches from other fields.

Vouros (Eds.), Coordination, Organizations, Institutions, and Norms in Agent Systems VI (Vol. 6541, pp. 172–191). Springer Berlin/Heidelberg. doi:10.1007/978-3-642-21268-0_10.

[216] Kaber, D. B. (2013). Adaptive Automation. In J. D. Lee & A. Kirlik (Eds.), *The Oxford Handbook of Cognitive Engineering*. Oxford University Press. doi: 10.1093/oxfordhb/ 9780199757183.013.004

[217] Kaber, D. B., Wright, M. C., Lawrence J. Prinzel, I., & Clamann, M. P. (2005). Adaptive Automation of Human-Machine System Information-Processing Functions. *Human Factors*, *47*(4), 730-741. doi: 10.1518/001872005775570989

[218] Parasuraman, R., Mouloua, M., Molloy, R., & Hilburn, B. (1993). Adaptive function allocation reduces performance costs of static automation. In *Proceedings of the Seventh International Symposium on Aviation Psychology*.

[219] van Diggelen et al., 'Designing for Meaningful Human Control in Military Human-Machine Teams'.

5. Conclusion and Future Research

This report, designated as version 2.0 of Deliverable in D2.1 within the ELSA Lab Defence project, investigated design methodologies aimed at tackling the ethical, legal and societal aspects (ELSA) associated with AI deployment in the military context. Based on literature study, the most important ELSA issues regarding AI were investigated, without particular focus on the use of AI in the military domain. A total of nine values affected by the use of AI were identified and described (dignity, human agency and autonomy, responsibility, life and physical integrity, privacy and data protection, liberty, justice, democratic decision-making and political participation, and peace and international security).

Next, existing ELSA design methods were identified and described. Most of these methods do not focus on defence and cannot directly be applied to the defence context, meaning that they may need to be adjusted and further tailored to military AI applications. A total of five major design methodologies are identified and described, all based on the concept of value sensitive design (VSD). These methodologies are: design for privacy/privacy by design (PbD), design for human agency and responsibility, including meaningful human control (MHC) and human oversight, explainable AI (XAI), and contestability by design, other approaches for design for human agency, including cognitive and socio-cognitive engineering, human-machine teaming and team design pattern engineering, engaging society in the design of military systems, including Scandinavian participatory design, critical design, speculative design, social design, and (new) participatory design, and design for emerging human-technology interactions, including coactive design and adaptive/adaptable automation.

These design methodologies are core design approaches and methods for mapping and addressing ELSA concerning new technologies. This provides an overview of the most relevant approaches and, together with the identified values affected by the use of AI, this allows applying these approaches to selected use cases in the military domain.

The design methodologies outlined in this report lay the foundation for the project's development of a comprehensive design methodology specifically tailored to AI in a military context. This customised methodology, which is possible due to its application to use cases, is intended to identify ELSA concerns in the utilisation of non-lethal AI technologies within the defence domain and offer guidance on crafting military AI technologies that pro-actively address ELSA-related challenges.