DELIVERABLE 2.1

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

Date	21 November 2023
Authors	Henning Lahmann, Bart Custers, Benjamyn I. Scott
Version	

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 AI in the military domain. The methodologies mapped in this report serve as the starting point for developing within the project a comprehensive design methodology tailored to identify ELSA issues in the use of AI technologies in the defence domain and provide guidance for designing military AI technologies that avoid or minimise ELSA issues. To map relevant technologies, a three-step approach is used.

First, based on a 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 six values affected by the use of AI are identified and described (dignity, privacy, life and physical integrity, liberty, democratic decision-making and political participation, and peace and international security). These six values, as located in the existing literature, are then put into the military AI context and linked to ELSA issues.

Second, 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 11 design methodologies are identified and described (value-sensitive design, guidance ethics, cognitive engineering, socio-cognitive engineering, coactive design, explainable ai, meaningful human control, team design pattern engineering, contestability-by-design, participatory design and evaluation methods, and privacy by design and privacy by default). These 11 design methodologies are core design approaches and methods for mapping ELSA concerning new technologies. This provides an overview of the most relevant approaches and allows them to be applied to selected use cases.

Third, to give the research real-world application, so that the results have usable benefits, the methodology is applied to 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 two established use cases are (1) Countering cognitive warfare using Early Warning Systems and (2) (Non-lethal) autonomous robots. Additionally, it was necessary to include a third use case (military decision-support system) as this allowed for additional ELSA to be demonstrated and described.

The introduction of new technology in defence offers opportunities, yet also creates risks. Introducing AI technology raises ethical, legal and social issues. If AI is to be applied 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

Chapter 1 - Introduction

1.1 The ELSA Defence Lab

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 present in the defence sector, there are 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 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.

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 on the ethical, legal and societal aspects 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 ethical, legal and societal aspects of military AI-based applications, including unmanned aircraft. It builds upon existing methods for value-sensitive design, explainable algorithms 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 several case studies. One of these case studies is on unmanned aircraft.[2] The ethical, legal and societal aspects 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]

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] 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.

1.2 The aim of this report

The use of AI in the military domain can cause ELS issues. This report aims to explore design methodologies to address these issues, but first, it is necessary to obtain a clear understanding of the issues. A lot of research already exists on the ELS issues of new technologies, such as big data technologies, data science and AI.[4] However, most of these issues are context-dependent and most of the existing research is not applied or tailored to the military domain.

The goal of our project is to develop an ELSA methodology for military AI applications. 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 issues. 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.

This chicken-and-egg problem is addressed by taking an iterative approach to our project. In this report, a first step is taken to achieve this. Based on examining existing literature, the most important ELSA issues are identified (Chapter 2). 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).

These steps (i.e., this report) constitute the first cycle in this iterative approach. The next cycles can be found in the subsequent report 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 literature study, complemented with online research. For mapping the most relevant values affected by AI (Chapter 2), we combined existing value lists of literature[5] on this topic and selected the values that are most relevant when focusing on AI use in the military. For mapping the most relevant design methodologies (Chapter 3), we combed through existing tools mentioned in literature in ethics, law, and social sciences. Through online searches, the list of design methodologies was verified and further expanded. In this case no selection was made, all approaches we could identify are listed and described in this report. As a result,

[4] 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] 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

some design methodologies may overlap: some approaches were developed in response to earlier approaches and some were developed to elaborate on earlier approaches. We list everything as it may all contribute to elements of an integrate model for the use of AI in the military.

1.4 The structure of this report

Chapter 2 examines the most important ELSA issues regarding AI that are mentioned and described in the existing literature. Most of this literature is not about AI applications in the military domain, but about AI in general or AI applied to other domains. As a first step towards the military domain, Chapter 2 provides examples of the values at stake and conflicting values when AI is applied in the military domain.

Chapter 3 provides an overview of the most important existing ELSA design 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 applied 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.

Chapter 4 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.

Chapter 2 - Affected values and potential value conflicts in the context of military AI

This chapter examines the most important ELSA issues regarding AI that are mentioned and described in the existing literature. The majority of existing literature is not about AI applications in the military domain, but about AI in general or AI applied to other domains (such as …?). It is beyond the scope of this report to describe all ELSA issues that AI can or could occur in different domains. Instead, this chapter identifies and describes the recurring values found in the literature. Furthermore, examples are provided of how these ELSA issues can or could play a role in military AI applications.

A short note on the term 'values' is required. The term 'values' is used to create a uniform approach towards ELSA. In all three areas (i.e., ethics, law and society), values (sometimes also referred to as 'ideals') describe the desirable state of affairs that one aims to realise. Principles and norms (for instance, ethical norms, legal rules, or societal conventions) prescribe how to achieve this state of affairs. For instance, justice is a value or ideal, as people generally agree that justice is important in a society. What the realisation of justice in society looks like, however, may not be known in detail.[6] People may disagree over what justice means, and according to which rules and principles it can be achieved. For instance, when it comes to a just distribution of scarce resources (i,e., distributive justice), for some, justice is achieved through equality (everyone getting an equal share), for others justice is achieved according to needs (bigger shares for those in need).

The term 'values' is useful in this context because ELSA are intertwined and sometimes hard to distinguish. When society establishes certain values, these will gradually turn into social norms. These social norms can also be framed as moral norms, as they refer to commonly accepted beliefs of standards of behaviour in a society. The moral part of these norms is the willingness to take the vulnerabilities of others into account.[7] When enforcement of these norms is important and social pressure alone is not sufficient, societies can codify these norms into law. Such legal norms can be enforced and even further developed via the judiciary, i.e., the court system.

When identifying issues with values, two types of issues are distinguished in this chapter:

- Type 1: violations of values
- Type 2: conflicting values

This chapter is structured according to the most relevant values in this context. Sections 2.1 through 2.6 respectively discuss dignity, privacy, life and physical integrity, liberty, democratic decision-making and political participation, and peace and international security.

2.1 Dignity

Human dignity is a value that sometimes seems to underlie all other values, particularly in human rights law and ethics. In ethics, dignity is the clearest way of taking the vulnerabilities of others into account, it is focused on the human condition, which is inherently vulnerable.[8] In human rights law, which focuses on the rights and freedoms of people, human dignity also seems to be the common

[6] See also Van der Burg (1996), p. 6.

[7] Frankena, W.K. (1973) *Ethics*, Englewood Cliffs, New Jersey: Prentice Hall.

[8] Fineman, M.A. (2008) The Vulnerable Subject: Anchoring Equality in the Human Condition. *Yale Journal of Law & Feminism*, Vol. 20, No. 1, 2008.

denominator. The German constitution starts with human dignity (*die Würde des Menschen*) as the first basic right listed. The same holds for the European Union (EU) Charter of Fundamental Rights, which starts in Article 1 with human dignity. In both legal documents, the right to human dignity is stated to be inviolable, it is an absolute right, meaning that under no circumstances violations of this right are allowed (contrary, for instance, to the right to privacy that can be overruled under circumstances, see Section 2.2).

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 or justice) are to some extent related to dignity. Interference with these values implies interference with human dignity. In other words, dignity is the underlying value or one of the underlying values that may need protection. The notion of dignity is very broad and subject to interpretation. For instance, for most people, it is clear that torture is a violation of human dignity.[9] Torture is often intended to humiliate people, which ipso facto affects their dignity. However, when it comes to personal freedoms relating to personal preferences and personal developments, views may differ. For instance, in some countries, freedom of religion or non-discrimination are seen as preconditions for allowing people to fully develop themselves and their lives, whereas in other countries these forms of pluralism are not considered dignified.

Dignity is under pressure in the context of AI.[10] In the information society, the reputation of people is increasingly constituted by the data that is 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) than as human beings of flesh and blood.[11] When a person no longer is treated as someone with particular interests, feelings and commitments, but merely as a bundle of data, that person's dignity may be compromised. The example of a black couple having been auto-tagged as gorillas,[12] also demonstrates this. Practices like profiling can reinforce a tendency to regard persons as mere objects.[13]

In the military domain, a typical example of this is the 'PlayStation mentality' in military drone use.[14] When drone pilots are deploying military drones 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 drones that kill real people. This can have negative and addictive effects.[15] It obfuscates the distinction between fiction and reality.

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

[10] Düwell, M. (2017) Human dignity and the ethics and regulation of technology. In R. Brownsword & E. Scotford (Eds.), *The oxford handbook of law, regulation and technology*. Oxford: Oxford Handbooks Online.

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

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

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

[14] Finn, R.L., & Wright, D. (2012). Unmanned aircraft systems: Surveillance, ethics and privacy in civil applications. *Computer Law & Security Review, 28*, 184-194.

[15] Blinka, L., & Mikuška, J. (2014). The role of social motivation and sociability of gamers in online game addiction. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, *8*(2), article 6. doi: 10.5817/CP2014-2-6.

AI can also put human dignity under pressure by treating them unfairly. A digital representation of a person can be different from reality in several ways. For instance, the data can be incorrect or incomplete which can create an incorrect representation. An example here is when people are suspected of planning a terrorist attack, but the names of the actual terrorists are identical to those of innocent people so the latter are being wrongly suspected to be terrorists. However, even if the data is correct, it can be weighed in different ways, stressing factors that weigh in less in real life. An example is when people are assessed based on their post code, which is rarely a criterion in real life. Further, even if the data and the weight of factors are correct, it can lead to unfair decisions. An example is when ethnicity or religion is used for decision-making, which violates anti-discrimination legislation in most jurisdictions.

Theories on value-sensitive design (VSD, see Chapter 3) do not refer to dignity but to identity. Yet, these two notions are intertwined.[16] In VSD, identity refers to people's understanding of who they are, embracing both continuity and discontinuity over time, but dignity also requires a similar understanding so that one can respect this value. Following from this it could be argued that dignity and identity mutually predestine each other: when dignity is preserved so is identity, and when identity is preserved so is dignity. Human dignity can, therefore, be regarded as a prime principle since all implications of digital technologies like AI affect humans and to differing degrees their identity and dignity.

The reduction of human beings to digital representations also affects their privacy, particularly their information privacy or right to personal data protection.

2.2 Privacy

Privacy as a value is often mentioned in the context of AI.[17] Privacy is a fundamental right (for instance, in the United Nations (UN) Universal Declaration of Human Rights, the EU Charter for Fundamental Rights, and in many national constitutions). The right to privacy originally focused on private and family life. In the information age, another aspect, i.e., informational privacy, has also gained importance. Informational privacy is closely related to the protection of personal data.[18] There are two major ways in which the privacy of people may be affected by developments in AI: violations of privacy may occur (1) when processing personal data, and (2) when AI tools disclose privacy-sensitive patterns.

To start with the first, there is an obvious 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 is processing. Without these large amounts of data, 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

[16] 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.

[17] Omer, T., & Polonetsky, J. (2012). Privacy in the age of big data: a time for big decisions. *Stanford Law Review Online, 64,* 63.

[18] 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.

to identified or identifiable natural persons). This is where personal data protection laws are relevant. In the EU, the General Data Protection Regulation (GDPR)[19] is an important legal framework regulating the collecting and processing of personal data.

The GDPR, which came into force in 2018, puts forward a set of principles and rules that data controllers need to take into account. First of all, all processing has to be lawful, fair, and transparent (Article 5(1) GDPR). Furthermore, the purposes for which the data are collected and processed have to be stated in advance (purpose specification) and the data may not be used for other purposes (purpose or use limitation) and data may only be collected and processed when necessary for these purposes (collection limitation or data minimisation). Data has to be accurate and up to date (data quality). When data is no longer necessary, it has to be deleted (storage limitation). The data needs to be processed in a way that ensures appropriate security and has to be protected against unlawful processing, accidental loss, destruction and damage (data integrity and confidentiality). Furthermore, the data controller is responsible for compliance with the law (accountability, Article 5(2) GDPR).

Within the scope of application of the GDPR, 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 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 (Article 9 GDPR). In case private actors process personal data, consent seems to be the most often used legal basis. There are many issues with informed consent in the context of AI, for instance, because it may be difficult to explain and understand how the AI tools are processing personal data.[20]

Data subjects have several rights regarding their personal data, including a right to transparent information on the data collected and the purposes for which it is processed (Articles 12–14), a right to access their data (Article 15), a right to rectification (Article 16), a right to erasure (Article 17), a right to data portability (Article 20), and a right not to be subject to automated decision-making (Article 22). For more details on data subject rights, we refer to the available literature.[21]

It is not only the use of personal data that can cause privacy issues. Also, what comes out of the processing of these personal data can be a privacy issue, particularly in cases in which the use of AI tools may disclose privacy-sensitive patterns. This can easily be seen when AI tools start prediction information that people may not want to disclose. 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. A typical example is tagging 'likes' on Facebook. Users can indicate, *inter alia*, with this icon what music, film clips, games, comments or people they like. Research shows that, based on only a few Facebook likes, it is possible to make a very accurate prediction of numerous

[19] https://eur-lex.europa.eu/eli/reg/2016/679/oj

[20] 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.

[21] Vrabec, H.U. (2021) *Data subject rights under the GDPR*, Oxford: Oxford University Press.

sensitive personal attributes.[22] For example, the researchers of Facebook users were able to make highly accurate predictions about their sexual preferences, ethnicity, religion, political preferences, personality traits, intelligence, happiness, drug use, and whether their parents were divorced. 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.[23] It should also be pointed out that anonymisation can also be reversed in the same way.[24]

The GDPR does not match well with the developments in AI.[25] 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. Furthermore, AI can to some extent deal well with inconsistent, incomplete or incorrect data. As long as there are large amounts of data available, AI tools can infer data and even correct errors in datasets. However, using flawed data is clearly at odds with the data accuracy principle that states that all personal data that is being processed should be accurate and up to date.

In the military domain, a lot of data that is processed is non-personal data that hardly affects privacy. For instance, surveillance data on landscapes, intelligence on a country's fleet, or flight routes of drones are unlikely to contain personal data. At the same time, however, in the military domain, there is also a lot of processing of personal data. Typical examples are interceptions of communications, the whereabouts of people participating in espionage activities, movements of troops of foreign nations, or information on distinguishing fighters from civilians. In the EU, the collecting and processing of much of this information is governed by national (military) intelligence acts rather than the GDPR. The GDPR has broad application and applies to all forms of personal data processing, unless exceptions apply, such as for the military or national security.

Note that the processing of personal data in the military domain is not restricted to personal data related to people working in the military domain. AI tools need to also process data from others to learn distinctions. For instance, to distinguish terrorists from non-terrorists, AI needs data on both categories of people. The same applies to distinguishing soldiers and combatants from civilians or to distinguishing spies from ordinary people. In other words, building profiles on these categories of people under scrutiny also requires (large amounts of) data on innocent, ordinary civilians.

[22] 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.

[23] 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/

[24] Ohm, P. (2010) Broken promises of privacy: Responding to the surprising failure of anonymization. *UCLA Law Review* 57: 1701

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

2.3 Life and physical integrity

Two values that inevitably gain significance in the context of the use of AI-supported systems in the military domain are 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.[26] It is explicitly acknowledged in all international human rights instruments and virtually all national constitutional arrangements. Its foundational character is exemplified in the text of the International Covenant on Civil and Political Rights (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). Crucially, 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 Covenant, this does not apply to the right to life (Article 4(2) ICCPR). Similarly, the European Convention on Human Rights (ECHR) provides that no one shall be deprived of his life intentionally (Article 2(1)). Even more succinctly, Article 2(1) of the Charter of Fundamental Rights of the European Union 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 by its very nature closely related to the right to life. While it is explicitly enumerated merely in the Charter of Fundamental Rights (Article 3(1)), 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.[27] 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.[28] Many national jurisdictions explicitly recognise the value as a constitutional right as well.

The extensive scope of the right to life, while it ought not to be interpreted narrowly,[29] 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 state.[30] 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

[26] UN Human Rights Committee, General Comment No. 36: Right to Life, 3 September 2019, UN Doc. CCPR/C/GC/36, at para. 2.

[27] 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.

[28] ECtHR, Case of X and Y v. The Netherlands, app. no. 8978/80, judgment, 26 March 1985, at para. 22.

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

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

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.[31]

In the general context of AI-supported systems, 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, i.e., whenever *safety* is a primary concern. This may be the case, for example, with autonomous or semi-autonomous robots that are employed in health or elderly care.[32] Another important class of use cases concerns the safety of autonomous vehicles like self-driving cars[33] or unmanned aircraft systems (UAS).[34] Here, the most important questions concern the establishment of adequate safety standards and other regulations that design organisations and manufacturers have to meet in order 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.[35]

Where AI-supported systems are employed in a military context, at least four different categories should be considered when it comes to life and physical integrity as fundamental values whose protection should be guaranteed.

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.

A second, more critical group of use cases involves general safety concerns not unlike those in the civilian contexts mentioned above, namely if military UAS navigate autonomously or the armed forces develop maritime or ground vehicles that are capable of operating without human intervention.

A third category 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[36] or large language models to generate subversive messages.[37] Given the content of the messaging, it cannot be ruled out *prima facie* that the cognitive influence leads to harmful effects, for example, in

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

[32] 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.

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

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

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

[36] 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.

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

the case that members from the target audience ingest poisonous substances due to disinformation.[38] 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.[39]

Finally, a fourth category of possible uses of AI-supported applications are of such a nature that the deprivation of life is the systems' 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 procedural 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.[40] To the extent that human rights law also envisages a justified deprivation of life in exceptional circumstances, as explained above, such conduct is likewise bound to strict procedural preconditions.

2.4 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 European Union.

In civilian security and criminal justice contexts, AI is increasingly being employed 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'),[41] which might lead to arrests and, thus, interfere with a suspected person's personal liberty. Other AI applications evaluate risks concerning

[38] 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.

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

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

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

decisions concerning bail,[42] sentencing[43] or parole.[44] 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.[45] 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.[46]

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 that deem them a security threat to the armed forces. To the extent that such decisions are already assisted by AI systems or will be in the future, the value of personal liberty must be considered when thinking about the broader implications of the use of AI in defence.

2.5 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 (Article 19 ICCPR; Article 10 ECHR; Article 11 Charter of Fundamental Rights), the freedoms of assembly and of association (Article 21 ICCPR; Article 11 ECHR; Article 12 Charter of Fundamental Rights), as well as the right to vote and to be elected, which is guaranteed by the ICCPR

[42] 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/.

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

[44] 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.

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

[46] 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.

(Article 25). In their collective manifestation, these values find their expression in the right to self-determination (Article 1 ICCPR).

Despite being mentioned less frequently in the context of the deployment of AI systems in the military, these values might gain relevance with regard to certain applications. Most strikingly, 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,[47] they nonetheless have important implications for these democratic values in general.[48]

2.6 Peace and international security

Peace and international security are frequently overlooked 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" (Article 1(1) UN Charter). 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.

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 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 armed UAS.[49] In light of the fact that this type of military operation vastly reduces the risk for one's 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.[50] 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.

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

[48] 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.

[49] 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.

[50] 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.

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.[51] 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) 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 (Article 51 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.[52]

[51] 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/.

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

Chapter 3 - Existing design approaches and methods for ELSA frameworks for AI applications

In this chapter, an inventory is offered of existing design approaches and methods for mapping ELSA of new technologies. Although much of the literature on ELSA issues related to AI and other technologies focuses on separate values such as human dignity, privacy, justice or autonomy, there is also literature offering more holistic approaches, allowing more comprehensive views and better balancing of multiple values. This chapter intends to offer an overview of the most relevant approaches in sections 3.1 through 3.11. We respectively discuss value-sensitive design (VSD), guidance ethics (GE), cognitive engineering, socio-cognitive engineering (SCE), coactive design, explainable AI (XAI), meaningful human control (MHC), team design pattern engineering, contestability-by-design, participatory design and evaluation methods, and Privacy by Design and Privacy by Default (PbD).

3.1 Value-sensitive design (VSD)

Technologies can be designed in different ways and such designs are never neutral in regard to (societal) values, interest, and rights. The way technology is designed usually depends mostly on functionality requirements, i.e., focusing on what the technology is supposed to do. For instance, face recognition tools should recognise faces. However, even from these specific functionality requirements, multiple designs can follow. Even when a technology can do what it was designed to do, there can be differences in, for instance, user friendliness or aesthetics of the design. Differences in design can also follow from different interpretations of functionality requirements, for instance, on what is a good performance. A design can be focused more on the speed of the task to be performed or more on the accuracy of the task to be performed. Some designs will focus on longer life expectancies of the technologies, whereas other designs may be focused on lower production costs. Some of these choices are made explicitly by those who design and manufacture these technologies, other choices may be more implicit, for instance, caused by convictions, beliefs, and behaviour of those who design new technologies. A typical example is that in the technology sector, there are more male workers than female workers. As a result, some technology designs are more oriented towards males than females. Take, for instance, facial recognition tools, which perform better on white, middle-aged males than other categories, because these tools were trained on datasets with white males.[53] Another example is the development and testing of new medicines, which often contain gender biases.[54]

Value sensitive design (VSD) is a theory that states that important values (like the values identified in Chapter 2 of this report) should be included in the design of new technologies. In other words, those designing and manufacturing new technologies should not only take into account functional design requirements, but also ELSA design requirements. This theory was developed by Batya Friedman and

[53] B. F. Klare, M. J. Burge, J. C. Klontz, R. W. V. Bruegge, and A. K. Jain (2012) Face recognition performance: Role of demographic information. IEEE Transactions on Information Forensics and Security, 7(6), p. 1789–1801. See also Fosch-Villaronga, E., Poulsen, A., Søraa, R.A., & Custers, B.H.M. (2021) Gendering Algorithms in Social Media. *ACM SIGKDD Explorations Newsletter*, 23(1), p. 24–31, https://doi.org/10.1145/3468507.3468512.

[54] Merone L, Tsey K, Russell D, Nagle C. (2022) Sex Inequalities in Medical Research: A Systematic Scoping Review of the Literature. Womens Health Rep (New Rochelle). 2022 Jan 31;3(1):49-59. doi: 10.1089/whr.2021.0083.

Peter Kahn at the University of Washington in the late 1980s and early 1990s.[55] As the name of this theory already indicates, it is a design-based approach, in which it tries to implement values into the design of new technologies, as opposed to an application-based approach, in which it tries to address how new technologies are used. As such, VSD must be applied before technologies are produced and put on the market.

VSD takes human values into account throughout the whole process of designing new technologies.[56] It requires that technological teams of designers and engineers are supplemented with other experts, such as experts in ethics, law and social sciences. The technology experts can establish functionality requirements and the other experts can establish the ELSA requirements. Of particular importance is that the ELSA experts find ways to translate the ELSA requirements into design requirements, which can be a difficult task. For instance, an ELSA requirement can be non-discrimination, but how is this translated into a design requirement? From an engineering perspective, non-discrimination concerning gender would mean that males and females are equally represented in models and that gender is not a decisive attribute in the modelling. However, for ELSA experts the focus would probably be more on the unfair treatment that could result from using gender as part of any modelling.[57] For instance, the use of gender is not illegitimate when used for affirmative action, but this could be hard to translate into the models.

When applying the VSD approach in practice, it usually consists of an iterative approach. The first step is a stakeholder analysis to get the most relevant perspectives on board. Next, with these stakeholders, the most important values can be identified and integrated into the design process. An initial design can then be assessed with the relevant stakeholders and adjusted and improved where needed. Although the theory has defined specific ways of doing this, in practice multiple variations of this approach exist.

Using a VSD approach is likely to yield technology designed in ways that better address ELSA issues. This can be done without loss of functionality and sometimes without extra costs. However, this may not always be the case, as some ELSA requirements may interfere with functionality requirements, which means these have to be balanced. Taking into account ELSA design requirements is likely to increase user acceptance and public support for new technologies, which can constitute a long-term interest for companies that develop new technologies. Apart from this interest, using VSD can also be a legal requirement in some situations. Typically, the GPDR requires that data controllers use data protection by design (DPbD) and data protection by default methods in some situations, see Section 3.11. In this context, typical examples of data protection by design are the use of encryption and anonymisation tools. A typical example of data protection by default is the use of opt-in versus opt-out.

[55] Friedman, B., Kahn, P. H., & Borning, A. (2006). Value sensitive design and information systems. In N. P. Zhang & D. Galletta (Eds.), *Human-computer interaction in management information systems: Foundations* (pp. 1–27). New York: M. E. Sharpe. Retrieved from https://cseweb.ucsd.edu/~goguen/courses/271/ friedman04.pdf.

[56] Friedman, Batya; Kahn, Peter H. Jr. (2002). "Value Sensitive Design: Theory and Methods". CiteSeerX 10.1.1.11.8020.

[57] Indre Zliobaite and Bart Custers, ‘Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models’, Artificial Intelligence and Law, 2016, 24:183-201.

3.2 Guidance ethics (GE)

Taking into account ELSA in the design and deployment of technologies is often considered a burden by those developing and using these technologies. They argue that taking ELSA issues into account will make things more complicated, take considerably more time and increase costs. Whereas the VSD approach starts from a kind of subsidiarity argument ("if two technologies can achieve the same goal, choose the one that is most ELSA compliant"), it can be criticised that two technologies are never equal, and this assessment usually involves trade-offs between functionality and ELSA issues. As a result, taking into account ELSA is considered a negative approach, in which there is too much focus on the negative aspects of new technologies. For instance, some argue that it is better to balance the weaknesses of AI with the strengths of AI. This can be broadened by also comparing the weaknesses and strengths of AI with those of humans.[58]

When comparing AI performance and human performance in both strengths and weaknesses, different kinds of arguments arise about whether to use AI. For instance, currently, there is a lot of critique of autonomous (self-driving) vehicles. All prototypes that are under development still sometimes cause accidents. Society seems to be waiting until the perfectly safe and reliable autonomous vehicle is manufactured (and then decides whether to use this technology). In reality, this may never happen, as no technology is that perfect. The tipping point is not when the technology is flawless in terms of functionality and safety, but when the technology outperforms humans. If, for instance, in a country, there are a thousand traffic casualties per year with human drivers and that could be reduced to a hundred casualties by replacing human drivers with autonomous vehicles, then this would be a convincing argument, even though the technology is not performing perfectly.

The GE approach was developed to open up narrow discussions on whether the use of specific technologies is ethical or not. The idea is to broaden the discussion on the pros and cons of a technology.[59] This approach tries to steer clear from the negative approach that technologies should not be used until all ELSA issues are resolved. The guidance ethics approach intends to be more constructive and give technological developments a desirable direction instead of merely fully embracing or rejecting them.

The guidance ethics approach does not consider ethics as a form of assessment, but as a tool to have a debate on ELSA issues. ELSA experts are not supposed to be external assessors, criticising technology developers, but should be right in the middle of the technology development together with the technology developers.[60] As such, the GE approach is design-based, just like VSD. The central question in guidance ethics is not binary ('yes or no'), but rather 'how to'. The core of the GE approach is to identify ELSA implications of technology and the central values at stake. This is done in a deliberative

[58] Meerveld, H.W., Lindelauf, R.H.A., Postma, E.O. *et al.* The irresponsibility of not using AI in the military. *Ethics Inf Technol* 25, 14 (2023). https://doi.org/10.1007/s10676-023-09683-0

[59] Verbeek, P.P., Tijink, D. (2020) Guidance ethics approach: ethical dialogue about technology with perspective on actions. Leidschendam: ECP. https://ris.utwente.nl/ws/portalfiles/portal/247401391/060_002_Boek_Guidance_ethics_approach_Digital_EN.pdf

[60] La Fors, K., & Meissner, F. (2022). Contesting border artificial intelligence: Applying the guidance-ethics approach as a responsible design lens. *Data & Policy*, *4*, e36.

process, in which concrete technologies are discussed instead of making a generic analysis of ELSA issues of technology.

It can be argued that the GE approach is not novel. In the end, other approaches, such as VSD, also focus on the how-question and deliberative processes. The guidance ethics approach mostly tries to address major points of critique on other existing approaches, which is that these are too generic and somewhat negative. The GE approach tries to address this by emphasising a more positive and concrete approach, but it remains to be seen whether this will have the intended effect.

3.3 Cognitive engineering

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.[61] 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.[62] Conceptually, it can be described as a type of applied cognitive science, in the sense that it builds on the discipline's findings to design and construct machines.[63] 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.[64] 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.[65]

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.[66] Specifically, Canellas and Haga have argued that "the foundation for a precise, comprehensive and robust definition of [meaningful human control] is found in the cognitive engineering discipline, whose primary focus is on the interaction between automation and humans, particularly in complex and dynamic domains".[67] Zooming in on the concept of 'function allocation' as developed within cognitive engineering, the Canellas and Haga contend that it provides an adequate theoretical framework for conceptualising

[61] 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.

[62] ibid.

[63] 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

[64] ibid.

[65] 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.

[66] Gersh, McKneely and Remington (n 53).

[67] 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.

the control of autonomous agents that are employed in public security scenarios. With reference to earlier work by Feigh and Pritchet,[68] they enumerate "five requirements for effective function allocation: (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."[69] This list of requirements then 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.[70] 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.[71]

3.4 Socio-cognitive engineering (SCE)

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.[72] 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.[73]

The 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.

3.5 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

[68] 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.

[69] Canellas and Haga (n 59).

[70] ibid.

[71] Canellas and others (n 57).

[72] 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.

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

account the ‘social’ aspects of working together.[74] Designing human-robot systems using traditional methods typically involves task allocation and decomposition. The best-known use of this strategy is supervisory control,[75] in which tasks are delegated to one or more machines, and their performance is subsequently observed. One of the problems with such methods 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.[76]

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.

3.6 Explainable AI (XAI)

In the field of AI, 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’.[77] In the last decade, the focus of this debate has shifted from transparency towards explainability.[78] 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 (see Section 2.2). Although for many types and applications of AI explainability may be

[74] https://dl.acm.org/doi/pdf/10.5898/JHRI.3.1.Johnson.

[75] Sheridan, T. B. (1992). Telerobotics, automation, and human supervisory control. Cambridge, MA: MIT Press.

[76] 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. 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.

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

[78] 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&nbsp;&hellip.

unnecessary, for critical applications, such as defence, it is essential for users to understand, trust, and manage the AI systems they deploy.[79]

The focus on developing explainable AI is intended as a step to move beyond incorporating values and principles into the design of AI and build mechanisms that demonstrate responsible behaviour. 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.[80] The XAI approach addresses this by concrete measures to increase understandability. This, in turn, can increase reliance, trust, and control concerning these systems. Similar to the VSD approach (see Section 3.1), the XAI approach focuses on implementing these measures in the technology design.

There are several ways in which AI models can be explained.[81] One way is by simplifying an AI system via approximation.[82] A second way is by explaining the features of an AI system.[83] A third way is by providing visualisations that are easier to understand for humans.[84] 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.[85]

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

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

[81] 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).

[82] 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.

[83] 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.

[84] 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.

[85] 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.

3.7 Meaningful human control (MHC)

Although definitions vary considerably,[86] the general notion of meaningful human control (MHC, also referred to as human-machine teaming) can be understood as a conceptual framework "to build and then utilise a partnership between people and AI so that each party utilises its strengths to achieve a shared goal".[87] 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.[88] 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.[89] 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.[90]

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.[91] Rejecting frequent suggestions that meaningful human control can be realised by simple adherence to the 'human-in-the-loop/human-on-the-loop' model, Van Diggelen *et al*. contend 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, etc."[92] 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 (see section 3.8). 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'.[93]

[86] J Christopher Brill and others, 'Navigating the Advent of Human-Machine Teaming', *Proceedings of the Human Factors and Ergonomics Society 2018 Annual Meeting* (2018).

[87] 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.

[88] Brill and others (n 78).

[89] ibid.

[90] 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.

[91] See Jurriaan van Diggelen and others, 'Designing for Meaningful Human Control in Military Human-Machine Teams' (arXiv, 12 May 2023) <http://arxiv.org/abs/2305.11892> accessed 14 June 2023.

[92] ibid.

[93] ibid.

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 teaming 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.[94] To enable an already constructed autonomous system to meaningfully interact with human operators as part of a team, Van der Vecht *et al*. propose 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).[95] In doing so, 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 states 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.[96] 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 understanding, predictability, familiarity, and context.[97] The Netherlands, too, used a variation of the MHC concept in its 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.[98]

3.8 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

[94] 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).

[95] ibid.

[96] UK Ministry of Defence, 'Joint Concept Note 1/18: Human-Machine Teaming' (2018).

[97] ibid.

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

common vocabulary and a set of well-defined solutions, facilitating communication among team members and promoting a shared understanding of the software architecture.[99]

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.

3.9 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".[100] As autonomous agents will remain brittle and unreliable, having not simply the right, but the technical-practical ability to contest predictive results of AI processes is critical to prevent harm.

In this context, it is important to realise that such contestation should not only come into play *ex post*, that is once the system has already produced a potentially injuring output. Instead, as Almada has pointed out, 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.[101] 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.[102] 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.[103] Thus, the approach is to be understood as a framework to design AI systems that enable

[99] 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.

[100] 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.

[101] 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.

[102] 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.

[103] Almada (n 93).

human intervention and actual challenging of decisions by involving human agency as early in the development process as possible, ideally through iterative stakeholder participation.[104]

To ensure contestability as the outcome of the design process of an AI system and not simply as a reactive right, Alfrink *et al*. have proposed a framework consisting of "five system features and six development practices".[105] The five system features are:

(1) built-in safeguards against harmful behaviour;
(2) interactive control over automated decisions;
(3) explanations of system behaviour;
(4) human review and intervention requests; and
(5) tools for scrutiny by subjects or third parties.

The six development practices are:

(1) ex-ante safeguards;
(2) agonistic approaches to machine learning development;
(3) quality assurance during development;
(4) quality assurance after deployment;
(5) risk mitigation strategies; and
(6) third-party oversight.

3.10 Participatory design and evaluation methods

Participatory design approaches build on the assumption that the introduction of potentially harmful novel technologies in value-sensitive environments should not be carried out without meaningfully involving as many affected stakeholder groups as possible. Among the principal motivating factors are an acknowledgement of information and power imbalances between those who design and employ AI systems on the one hand and impacted individuals on the other, as well as an urge to take adequate account of the most vulnerable.[106]

Regarding the aforementioned method of contestability-by-design (Section 3.9), Alfrink *et al*. have maintained that given the need to involve stakeholders at the earliest stages of designing an AI system, participatory methods present the most obvious and most suitable way to go about this. The clearest advantage is that any foreseeable issues with an autonomous agent, for instance, regarding a possible violation of fundamental rights, can be recognised and addressed before any harm is done. In order for this to happen, it is crucial that those stakeholders are involved who will most likely be negatively affected, even if it will usually not be feasible to let all potentially impacted individuals actively take part themselves. Finding the right representatives from relevant groups to ensure the appreciation of all interests is, therefore, one of the core tasks in participatory design processes. Furthermore, participation should not be limited to one-off consultations but consist of ongoing dialogues on an

[104] See next section, "Participatory Design and Evaluation Methods".

[105] Alfrink and others (n 92).

[106] Janet Davis, 'Design Methods for Ethical Persuasive Computing', *Proceedings of the 4th International Conference on Persuasive Technology* (Association for Computing Machinery 2009) <https://doi.org/10.1145/1541948.1541957> accessed 15 June 2023.

equal footing that clearly communicate to participating stakeholders that their insights, opinions, and arguments are not just valuable and legitimate by themselves but also taken into serious consideration when finalising the design process of the AI system.[107]

Many participatory processes are conceivable, from the setting up of focus groups to the direct involvement of selected individuals in the design process. Within the framework of the ELSA Lab Defence, one such instantiation of stakeholder participation concerning the development and deployment of AI-supported systems in military technology was a workshop with about 30 participants conducted during the 2023 REAIM conference in the Netherlands[108] that made use of the BetterBeliefs platform.[109] The developers describe their software as "a Bayesian social platform for inclusive and evidence-based decision making"[110] to utilise collective intelligence in value-sensitive settings.[111] The platform was first deployed during a workshop in Canberra in 2019 conducted by the Australian Department of Defence to discuss questions surrounding the development of ethical AI for use in military applications.[112]

Despite the relatively small number of participants, the explorative setup without dedicated participant preparation, and the rather limited timeframe of one hundred minutes of running time, the workshop did expose some of the opportunities and challenges of participatory design methods.[113] Thus, it became clear that while individuals from the public seem interested and eager to engage with difficult ethical, legal, and social questions concerning the increasing use of AI in security settings and the military in particular, any meaningful deliberation requires a significant amount of preparation, resources, and time commitment to produce outcomes that can usefully inform concrete development processes in the realm of technology as complex as AI.

3.11 Privacy by design and privacy by default (PbD)

Some of the approaches discussed in the previous sections are mandatory from a legal perspective. Most notably the VSD approach (discussed in Section 3.1) is, in a way, a legal obligation under the EU GDPR. 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

[107] Alfrink and others (n 92).

[108] https://reaim2023.org/.

[109] ELSA Lab Defence, 'Applying participatory methods to design complex military technologies' (*ELSA Lab Defence*, 27 February 2023) <https://elsalabdefence.nl/applying-participatory-methods-to-design-complex-military-technologies/> accessed 15 June 2023.

[110] Susannah Kate Devitt and others, 'A Bayesian Social Platform for Inclusive and Evidence-Based Decision Making' (arXiv, 13 February 2021) <http://arxiv.org/abs/2102.06893> accessed 15 June 2023.

[111] See https://betterbeliefs.com.au/index.php/solutions/.

[112] Kate Devitt and others, 'A Method for Ethical AI in Defence' (Australian Government Department of Defence 2020).

[113] Marc Steen, Jurriaan van Diggelen and Kate Devitt, 'Experiences in Participatory Design Using the BetterBeliefs Tool During the REAIM 2023 Conference' (ELSA Lab Defence 2023).

necessary safeguards into the data processing. This approach is called data protection by design (DPbD) or simply privacy by design (PbD).[114]

In essence, 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 (usually no more than 5 to 10 %) change default settings, privacy by default can be an important tool to preserve and protect privacy.

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. And 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.

Section	Approach	ELSA focus
3.1	Value-sensitive design (VSD)	Ethical/Legal
3.2	Guidance ethics (GE)	Ethical
3.3	Cognitive engineering	Ethical
3.4	Socio-cognitive engineering (SCE)	Social

[114] 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.

3.5	Coactive design	Social
3.6	Explainable AI (XAI)	Ethical/Social
3.7	Meaningful human control (MHC)	Ethical/Social
3.8	Team design pattern engineering	Social
3.9	Contestability-by-design	Legal
3.10	Participatory design and evaluation methods	Ethical/Social
3.11	Privacy by Design and Privacy by Default (PbD)	Legal

Table 3.1: Existing design approaches and methods for ELSA frameworks for AI applications.

Chapter 4 - Critical evaluation of design approaches and methods in view of domain specificities

This Chapter aims to provide a preliminary evaluation of the listed design approaches in light of the specific context of AI-supported systems in the military domain and the likely affected values as detailed in Chapter 2. Recognising the military domain as an especially ‘morally sensitive context’,[115] it can be expected that the most appropriate approach to address ELSA issues is to a large extent dependent on the specific use case, as the specific context and circumstances of a case determine what values will presumably be affected by the employment of the respective AI technology.

When it comes to autonomous weapons systems, the most immediately affected values are the dignity, life and physical integrity of persons who are targeted by the system or are in the vicinity of such a system during engagement. Such a highly precarious context requires particularly well-calibrated considerations when it comes to design approaches, as exemplified in both academic scholarship and State practice. For example, Boshuijzen-van Burken has demonstrated the challenges of applying the value-sensitive design approach to autonomous weapons. One major obstacle to overcome in this context, the author suggests, is the classified nature of the design process, which makes it inherently difficult to involve other stakeholders such as the general public.[116] Without making those frameworks explicit, in its ‘Directive 3000.09: Autonomy in Weapon Systems’, the US Department of Defense took into account insights from cognitive design, explainable AI and human-machine teaming. In this directive, for example, such systems are required to “be readily understandable to trained operators, such as by indicating what actions operators need to perform and which actions the system will perform”; “provide transparent feedback on system status”; “provide clear procedures for trained operators to activate and deactivate system functions”; and that

[115] Marc Steen and others, ‘Meaningful Human Control of Drones: Exploring Human–Machine Teaming, Informed by Four Different Ethical Perspectives’ [2022] AI and Ethics <https://doi.org/10.1007/s43681-022-00168-2> accessed 19 January 2023.

[116] Christine Boshuijzen-van Burken, ‘Value Sensitive Design for Autonomous Weapon Systems – a Primer’ (2023) 25 Ethics and Information Technology 11.

the physical hardware and software will be designed with appropriate "human-machine interfaces and controls"; or "technologies and data sources that are transparent to, auditable by, and explainable by relevant personnel".[117] Even though these cases concern lethal autonomous weapon systems and are, thus, not directly relevant to the ELSA Lab Defence, they may serve as expedient illustrations of the larger issues at hand.

This chapter sketches the specificities of the military domain to identify what is different when fielding AI-supported systems in defence as opposed to in a purely civilian context by way of investigating the use cases. The research question to be answered is, in a tentative fashion, is: which existing design approaches are most suited to accommodate AI uses in the military domain?

Section 4.1 describes the two use cases selected by the ELSA Lab Defence consortium to date and adds a third one that appears expedient to sketch some of the critical value conflicts in the context of military AI applications. Section 4.2 investigates the values affected in each use case. Section 4.3 provides a preliminary discussion of (applicability of) the design approaches listed in Chapter 3 to AI in the military domain.

4.1 ELSA Lab Defence use cases

Two 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), and
(2) (non-lethal) autonomous robots.[118]

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.

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, which allows the 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

[117] Office of the Under Secretary of Defense for Policy, 'DOD DIRECTIVE 3000.09: Autonomy in Weapon Systems' (US Department of Defense 2023) <https://www.esd.whs.mil/portals/54/documents/dd/issuances/dodd/300009p.pdf>.

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

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 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 otherwise 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.

During the project, more case studies may be selected within the consortium, for example, from the cyber, maritime or space domain. The selection and identification of relevant military applications of AI for which ELSA are important is part of the project and will be done via the technology monitor.

For this initial mapping exercise, however, it has been deemed necessary to add a third preliminary use case to be better able to demonstrate and describe some of the values that are potentially impacted by the use of AI-supported systems in military applications.

military decision-support system

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 drones, 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.

4.2 Potentially impacted values

As laid out in Chapter 2, the use of AI-supported systems in defence contexts will inevitably create ELSA challenges in the form of impacted values. This also applies to the three use cases described in section 4.1, the use of AI against cognitive warfare, the deployment of non-lethal autonomous robots in military missions, and the use of AI-enabled military decision-support systems. Before providing a preliminary assessment of suitable design approaches, it is necessary to briefly explicate which of the values identified will likely be specifically relevant in these three contexts. As will become clear from the following analysis, not all of the listed values will be relevant in every use case; and even within a single devised scenario, not all values will be affected automatically, at all times or regarding all possible modes of operation.

Use case 1: countering cognitive warfare using Early Warning Systems

Concerning the use of a cognitive warfare monitoring and alert system, the first value potentially affected is the right to privacy and data protection. As explained above, how AI-based systems may infringe upon privacy is twofold. First, when personal data are processed as input for machine-learning processes, either for training purposes or during deployment; second, when the algorithmic output discloses privacy-sensitive patterns. Any evaluation will be incomplete without knowing the exact details of the system’s mode of operation, in particular, details of the data sources tapped for the monitoring and alert functions to work as intended. Privacy concerns may be relevant depending on the circumstances. As identified from past instances of foreign influence campaigns, for example, the Russian targeting of British and American citizens ahead of the 2016 Brexit referendum and the US presidential elections, contemporary cognitive warfare exploits personality traits as 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.[119] In such a scenario, detecting and countering this external threat might require the collection and analysis of one'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.

Other values are the ones pertaining to democratic decision-making and political participation. 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 Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression pointed out in regard to 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".[120] 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".[121] 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.

Use case 2: autonomous robots

In light of the large variety of possible applications of autonomous robots in the military context, possible impacts on the values identified in Chapter 2 are less straightforward to pinpoint. Nonetheless, a few more speculative considerations follow.

First, if the robot is equipped with sensors to observe its environment, collect audio-visual data and autonomously analyse the data about possible threats or targets by means of installed software, it is again privacy concerns come into play. This is particularly the case if the robot is fielded in urban

[119] 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.

[120] UN Doc. A/HRC/44/49, 23 April 2020, at para. 42.

[121] ECtHR, Salov v. Ukraine, 2005, at para. 113.

surroundings in which a large number of civilians are present. Surveilling their movements and quotidian activities quite obviously has considerable privacy implications.

Second, in case the autonomous robot gathers the data for the purpose of detecting and determining possible targets for an armed engagement by (human) forces, for instance, through an airstrike, then its activities have at least an indirect impact on the right to life and to the physical integrity of all persons present in the vicinity of the autonomous agent. The latter value can also be affected more immediately: if the scenario concerns a ground-based robot (as opposed to a UAS, for example) that moves around autonomously, then there are safety concerns to be taken into account that are reminiscent of challenges encountered in the context of self-driving cars. That means that these robots must be capable of reliably identifying obstacles, in particular any persons present in its 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.

Third, 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 anxiety.[122] 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.

Use case 3: military decision-support systems

In the case of military decision-support systems, again much depends on the concrete scenario and purpose for which such systems are being deployed. One of the principal tasks carried out by such an application is conducting ISR combined with an automated analysis of the gathered data that could lead to lethal targeting decisions in a situation of armed conflict. In such situations, the values most directly affected are the life and physical integrity not only of the person(s) targeted but also of any incidentally present bystanders, for instance, otherwise uninvolved civilians in an urban conflict environment.[123] The recommendation of a military decision-support system might produce (i.e., its predictive outcome as the result of data analysis) does not necessarily have to be an armed

[122] 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>.

[123] 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/>.

engagement with an identified target. Instead, the algorithmic system might recommend taking a person who has been deemed a threat into preventative detention.[124] 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.[125] Such conduct based on recommendations made by the decision-support system would have a direct impact on the liberty of the concerned persons.

Apart from tactical decisions on the battlefield, AI-enabled decision support systems have also been envisioned for higher-level strategy recommendations that concern the question of whether to resort to the use of force against an adversary in the first place.[126] Such systems would be tasked with analysing data for example from satellites or UAS to detect troop movements, the construction of military objects close to international borders, fused with other intelligence such as suspicious business transactions, conversations among the adversary’s officials and diplomats both on open platforms and in private, and other relevant data points. Proponents of these technologies expect them to deliver real-time insights into an adversarial actor’s actions and intentions that the algorithmic system can seamlessly translate into predictive output that increases awareness of the geopolitical situation and actionable intelligence as to whether armed forces should be activated and force be employed with the aim of thwarting the emerging threat.[127]

Whether each of the described versions of an AI-enabled military decision-support system is functional entirely depends on the constant collection of massive amounts of accurate, timely, diverse, nuanced, and above all constant-specific data, both to train the models before deployment and as input data to generate predictions and recommendations. As all conceivable scenarios concern areas of application that involve the presence of a civilian population, it is inevitable that the private data of civilians will be among the collected information used to run the algorithms. Therefore, the value of privacy will almost certainly be affected by the use of such systems.

Such constant surveillance practices to collect relevant data about a conflict or other target area have broader implications. If the surveillance is carried out by drones, among other assets, it is reasonable to expect effects as already laid out in the context of use case 2: If at the end of these algorithmic data practices stands a potentially lethal targeting decision, the continuous presence of drones will likely cause states of anxiety within the civilian population that directly impacts their ability to live a free and self-determined life and thus their dignity. In such a scenario, empirical findings show that an additional likely consequence of such military activities is a gradual decline in communal political

[124] 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/>.

[125] 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>.

[126] Ashley Deeks, Noam Lubell and Daragh Murray, ‘Machine Learning, Artificial Intelligence, and the Use of Force by States (2019) 10 Journal of National Security Law & Policy 1.

[127] See for one example of such a system, offered by the private company Rhombus Power, Stanley McChrystal and Anshu Roy, ‘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/>.

activities among civilian populations under such surveillance due to perceived risks of harm from algorithmic targeting decisions.[128] It, thus, seems justified to also consider the value of democratic decision-making and political participation potentially affected, even if in a completely different way as compared to use case 1.

Table 4.1 provides an overview of the values affected in each of the three use cases.

	Use case 1	Use case 2	Use case 3
Dignity		X	X
Privacy	X	X	X
Life and physical integrity		X	X
Liberty			X
Democratic decision-making and political participation	X		X
Peace and international security			X

Table 4.1: Overview of the value affected in each use case

4.3 Preliminary discussion of design approaches to AI in the military domain

Having identified the values most likely affected by the use of AI-based systems in the two use cases, this section provides a preliminary attempt to assess the various design approaches introduced in Chapter 3 in light of the particular challenges of the military domain. We examine whether there are, at this stage of the project, certain approaches – or combinations thereof – that appear to be more suitable than others and, thus, deserve a more detailed inquiry.

As was already hinted at, the different methods should not be considered necessarily mutually exclusive. Some approaches can indeed only come to fruition if they are applied to complement one another. For instance, value-sensitive design envisions a participatory stakeholder process that should be informed by insights from other participatory design and evaluation methods as described in section 3.10. Likewise, it has been pointed out that contestability-by-design should be implemented through early input from relevant stakeholders.[129] The latter approach can furthermore only be realised if the (potentially) affected person has sufficient and meaningful information about the algorithmic process, which could be implemented through explainable AI.[130] Methods that try to design AI systems that can account for ethical frameworks, such as value-sensitive design or guidance ethics, only make sense if the operator of the system is capable of acting accordingly, which requires taking on insights from cognitive engineering or human-machine teaming. Quite obviously, merely having a human 'in' or 'on the loop' will not suffice to actualise ethical standards if these humans do

[128] See e.g. Alex Edney-Browne, 'The Psychosocial Effects of Drone Violence: Social Isolation, Self-Objectification, and Depoliticization' (2019) 40 Political Psychology 1341.

[129] Almada (n 93).

[130] ibid.

not have the authority or technical ability to effectively intervene in a timely manner.[131] In this respect, Steen *et al*. laid out a productive case study to show the interplay between ethical guidelines and human-machine teaming in UAS.[132]

That being said, the military domain does present some specific challenges that do not always easily map onto existing design approaches. These challenges can broadly be put into four preliminary categories:

(1) applicable legal framework;
(2) applicable ethical framework;
(3) stakeholder participation; and
(4) criticality and time-sensitivity.

Here, we discuss these four categories of challenges.

First, approaches that rely on the applicability of certain legal frameworks that have a particular geographic or material scope of application (instead of relying on universal rules or general principles) might not easily be implemented in every conceivable military use case. For example, the GDPR will usually not be applicable when armed forces from an EU country act, in particular not if they are deployed outside the EU, which might hamper the usefulness of the privacy-by-design approach. Even if it is agreed to be applied in spirit, military leaders might not feel bound to the same extent.

Second, trying to make use of value-sensitive design or guidance ethics in the context of military AI raises the question of whose ethical frameworks should count and be decisive. Even in an entirely Western framing, many different norms and guidelines are conceivable. However, considering the armed conflicts with EU or NATO involvement over the past two decades, deployment of a military AI system in an entirely disparate setting is highly likely. In light of this, it is questionable whether Western value propositions should simply be imposed on potentially affected civilian populations with distinct cultural backgrounds.

Third, this reflection leads to the challenge of stakeholder participation. The leading value-driven approaches factor in participatory deliberative processes to stake out competing ethical claims. Meaningful stakeholder participation is already a considerable task in domestic contexts.[133] If we understand the notion of stakeholders as including representatives from the groups most likely affected by the AI system, this could call for the involvement of persons from populations in which, for example, an autonomous surveillance robot or UAS will be fielded. It is completely unclear how this could be arranged meaningfully, or why this should not be necessary at all. To point to one pertinent example in this context, when the US Department of Defense published its 'Ethical Principles for Artificial Intelligence' in February 2020, it praised a 15-month-long process of 'consultation with leading AI experts in commercial industry, government, academia and the American public".[134]

[131] Alfrink and others (n 92).

[132] Steen and others (n 107).

[133] Wessel Reijers and others, 'Methods for Practising Ethics in Research and Innovation: A Literature Review, Critical Analysis and Recommendations' (2018) 24 Science and Engineering Ethics 1437.

[134] US Department of Defense, DoD Adopts Ethical Principles for Artificial Intelligence, 24 February 2020, https://www.defense.gov/News/Releases/Release/Article/2091996/dod-adopts-ethical-principles-for-artificial-intelligence/.

However, all these are most likely not the people who will eventually be confronted with AI-based systems.

A further problem for participatory design processes is the culture of default secrecy when it comes to military technology. The classified nature of most development efforts makes it difficult to generate useful and sufficiently nuanced engagement from stakeholders who are not privy to the technical details of a system or its actual capacities. Boshuijzen-van Boken suggests that one workaround is the discussion of merely hypothetical and relatively high-level, abstract use cases.[135] While this might yield expedient results concerning some ethical or other relevant issues, it remains questionable whether participatory processes operating behind a veil of ignorance can ever get to the most critical and contentious matters that would require the most granular engagement.

Fourth, most military use cases are by definition particularly high-stakes and time-sensitive, as already pointed out in the previous section. This might pose a problem for approaches that revolve around the possibility of rectifying detrimental results of algorithmic processes. Clearly, 'contestability' as a concept is potentially problematic in the context of military AI at least about such decisions or predictions that may lead to harmful or even lethal outcomes, even if this concerns merely an autonomous surveillance robot rather than lethal autonomous weapons. This consideration remains relevant if we take into account that the contestability-by-design approach is not merely conceived as a purely *ex post* remedy but captures *ex ante* intervention and contestation. Accordingly, in introducing their contestability framework, Alfrink *et al*. caution that it precludes 'extreme high stakes contexts' and that it assumes situations "where time-sensitivity of human intervention is relatively low".[136] To be sure, this does not per se foreclose the method's expediency concerning all possible military use cases, but its limitation in this regard should be kept in mind.

[135] Boshuijzen-van Burken (n 108).

[136] Alfrink and others (n 92).

Chapter 5 - Conclusions

This report, designated as the initial Deliverable in WP2 (D2.1) within the ELSA Lab Defence project, seeks to investigate design methodologies aimed at tackling the ethical, legal and societal aspects (ELSA) associated with AI deployment in the military context. The methodologies outlined in this report lay the foundation for the project's development of a comprehensive design methodology. 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 mitigate or minimise ELSA-related challenges.

Both fundamental and applied scientific research are deployed, particularly through experimental methods, to establish a methodology aimed at translating Ethical, Legal, and Societal Aspects (ELSA) into socio-technical requirements. This methodology also lays the foundation for a practical and pragmatic research, development and deployment process for AI-based applications. It is rooted in a profound understanding of various aspects, including value-sensitive design, ethics, legal frameworks, public perception, team composition, and socio-cognitive engineering. Furthermore, this methodology extends its applicability to military AI-based systems.

Within this methodology, specific design patterns are outlined to enhance the alignment of targeted AI-based systems with ELSA considerations. These patterns address important aspects such as explainability, trust calibration, system state awareness, decision control, and collaborative learning in human-machine teams. Additionally, the methodology provides guidance on programming and calibrating AI components and includes system-level algorithms designed to expedite the implementation of ELSA principles.

D2.1 is just the first stage in Work Package 2.

- D2.2 will build upon this research. This will be because it will be dedicated to crafting a thorough ELSA impact assessment methodology. This methodology is designed to enable the analysis of military AI-based applications from various ELSA dimensions, ensuring a comprehensive evaluation of their ethical, legal, and societal impacts.
- D2.3 is dedicated to formulating design and development patterns tailored for military AI-based applications. These patterns are intended to provide guidance for creating applications that align with both current and foreseeable ELSA values.
- D2.4 is focused on developing reusable algorithms and system components. These components are designed to be incorporated into AI-based military applications to effectively address ELSA requirements.
- D2.5 focuses on the synthesis and operationalisation of the outcomes from WP2. This task aims to combine the results into a comprehensive methodology that can effectively ensure the alignment of military AI-based applications with both existing and future ELSA values.

WP2 serves as the central tool utilised for both stakeholder and public evaluation in WP3 and the case studies in WP4. The ELSA methodology encapsulates the research findings in a practical and actionable format for stakeholders.

END