2 A Socio-Technical Feedback Loop for Responsible Military AI Life-Cycles from Governance to Operation

Marlijn Heijnen, Tjeerd Schoonderwoerd, Mark Neerincx, Jasper van der Waa, Leon Kester, Jurriaan van Diggelen, and Pieter Elands

INTRODUCTION

Major investments in AI technologies are extending AI capabilities substantially, leading to new applications in the military domain. Such applications often contain embedded learning algorithms and some form of autonomous information processing and decision-making. These applications are highly needed in complex and time-critical military operations, such as AI-based cybersecurity countermeasures in response to adversarial (AI-based) attacks, or deployments of autonomous weapons for a ship's

DOI: 10.1201/9781003410379-3

self-defense. In general, AI can substantially reduce the risks and improve the precision of military operations, due to its capacity to process large amounts of data quickly and by reacting quickly based on the learned and pre-programmed models.

However, the military application of AI is under debate. For example, scientists and non-governmental organizations have warned against the emergence of "killer robots",[1] that is, autonomous weapon systems that select and attack targets without meaningful human control (MHC). Generally MHC refers to the requirement that not AI but humans should ultimately remain in control of and morally responsible for (AI-driven) military operations. However, in its advice to the Dutch government, the Advisory Council on International Affairs/Advisory Council on Peace and Security (AIV/CAVV) concluded that there is not yet agreement on an international definition of MHC, except that human judgment should always be preserved. In a philosophical account of MHC, Santoni de Sio and Van den Hoven (2018) identified two general necessary conditions for MHC: tracking (a system should be able to respond to the moral reasons of humans deploying the system) and tracing (a system should always allow to trace back the outcome of its operations to at least one human along the chain of design and operations).

The concern about MHC is not without reason: AI technology might bring along unintended (side-) effects that must be prevented, such as deterioration of human control, an unbalanced ("biased") situation awareness (i.e., caused by the frequency distribution of objects or phenomena in the training dataset), or the opponent or enemy anticipating predictable behavior of AI technology. A multitude of, often interdependent, factors can bring about such unforeseen negative effects, such as shortcomings in the technology (e.g., biased training data, incomplete world models, poorly designed user interfaces), and performance changes of humans who work with the AI systems (neglect due to loss of oversight or over-reliance). The challenge is to establish MHC: enabling humans to have oversight and take responsibility in decision-making (Aliman, 2020; Amoroso & Tamburrini, 2020; Scharre, 2018; Van Diggelen et al., 2023) throughout the AI lifecycle: MHC is not only to be achieved during the operation of AI-based systems, but also during governance, design and development activities.

To study responsible AI, we must set it in a realistic context. This can be done using scenarios that are operationally relevant, capture the complexity of defense operations, and express a clear need to use AI. In this chapter, we use a short scenario described in Box 2.1 to illustrate that moral decisions regarding the deployment and use of AI systems are made at several stages. The combination of these decisions determines how the use of the AI system is embedded and controlled in the operation, how the uncertainties are taken into account, how the risk assessments are made, and how the responsibilities are allocated (cf. Ekelhof, 2019). In this example, the military organization decided to use AI-enabled and remotely-activated, rapid-fire guns, and the commander authorized the installation and use of these guns to guard against light vehicles with explosives within a demarcated defense zone. When the risk of "vehicle attacks" is deemed high, a human guard can command the guns to fire at identified hostile vehicles. In this scenario, there are risks of collateral damage and incorrect target identifications.

Note that Box 2.1 presents a small and simplified scenario, in which, for example, changes of the human-machine capabilities are not addressed. Incidents and problems can appear and accumulate at the individual entity (e.g., malfunctioning actuators),

team organization (e.g., inappropriate allocation of responsibilities), and societal level (e.g., discrimination against specific population groups). As the technology is new and adaptive, and is operating in a dynamic environment, there will be uncertainties in the predictions of outcomes. The military decision-making processes, embedding advanced AI technologies, should incorporate careful consideration and weighing of the relevant ethical, legal, and societal aspects (ELSA), taking into account the uncertainties, risks, and unintended side effects. For example, a diminished value awareness can be a risk; we might think that an AI system is aligned with our values, while in reality they are only partially aligned (such circumstances set high requirements for AI technology's explanation capability).

In the example, objectives include the prevention of further violent escalations and supporting the government's administration, rule of law, and law enforcement. It represents a non-international armed conflict, to which Common Article 3 of the Geneva Conventions applies.[2]

What are the challenges concerning the design, development, and maintenance of human-AI systems, and how to identify the moral consequences of the deployment of

BOX 2.1 SIMPLIFIED MILITARY SCENARIO TO ILLUSTRATE MORAL DECISION-MAKING AT THE LEVEL OF AVAILABILITY, DEPLOYMENT, AND USE OF A (SEMI-) AUTONOMOUS AI SYSTEM[3]

RAPID DEFENCE SCENARIO

An urban operating base has been under attack by terrorists for several months. These attacks largely involve automobiles, disguised as civilian traffic but equipped with large quantities of explosives, driven by suicidal adversaries who accelerate when nearing the entry gate to the base. Since buildings and base personnel are located near the gate, there is very limited time for gate guards to target and respond to such attacks even when they can identify them, and much destruction and death have resulted over the past months.

To improve reaction times, the base commander previously authorized the installation and use of *RivalReveal*, that is, remotely activated, rapid-fire guns capable of stopping a light vehicle as it heads toward the gates. These guns can fire at various levels of autonomy. They can fire "automatically" within a previously defined target zone, after receiving prior orders from a human guard. Due to space limitations in the urban environment, the presence of base personnel in the target zone cannot be completely avoided. This means that there is a possibility that guns cause collateral damage to innocent bystanders. This risk of collateral damage to innocent bystanders is the primary hazard. Another risk is that the human guard will incorrectly identify a target, either positively or negatively. This scenario is suitable for following the Socio-Technical Feedback (SOTEF) loop, to be described below, because the violent nature makes it highly morally sensitive, and the high-speed decision-making justifies the choice of considering AI-based techniques.

new AI technologies in future operations? This is challenging, in particular, because these questions need to be addressed continuously during the complete lifecycle of the socio-technical systems[4] (STS), while the operational circumstances and conditions are continuously changing, affecting the decision-making processes and outcomes. Furthermore, the values concerning certain military operations can change over time. This means that the decision-making processes and outcomes should continuously be evaluated to ensure alignment with values. Thus, *value alignment* is an important continuous process for the identification of the moral values that are at stake. However, it remains difficult to identify relevant moral values (especially considering that they may change over time), and to ensure that the human-AI system continues to operate in accordance with these moral values and their context-dependencies. Another difficulty lies in ensuring that the human-AI system can notice in time when an outcome is in violation of one or more moral values. This is especially relevant for military AI systems, as a violation of values might have a severe impact.

To date, there is no consensus on how MHC must be operationalized (AIV/CAVV, 2021)[5] and how to achieve value alignment in the development and deployment of AI. There is agreement on guiding principles, such as formulated by the UN Group of Governmental Experts,[6] NATO,[7] and the TAILOR consortium.[8] The NATO Principles of Responsible Use for AI in Defense will help steer efforts in accordance with moral values, norms, and international law, but a comprehensive prescriptive approach is lacking for building and implementing AI technology in such a way that it is under MHC, during its complete lifecycle. In this chapter we propose the SOTEF loop: a methodology to establish MHC at the levels of society, organization, and operation, addressing regulation, design, development, maintenance, and modification processes of a specific human-AI system in a specific context (Aliman et al., 2019; Aliman & Kester, 2022; Peeters et al., 2021).

Known approaches such as value-sensitive design (Friedman & Kahn, 2003; Friedman & Hendry, 2019; Van Den Hoven, 2013), participatory multistakeholder analyses of the ELSA[9] (Van Veenstra et al., 2021), and responsible research and innovation (Stilgoe et al., 2013; Von Schomberg & Hankins 2019),[10] share important aspects with the SOTEF loop such as stakeholder involvement, value analyses, and multidisciplinary design. Other guidelines (e.g., Dunnmon et al., 2021) also have become more concrete on how to implement ethical principles. The SOTEF loop incorporates these approaches and applies them not only in the design phase of an AI system, but ensures value alignment throughout the STS lifecycle. The SOTEF loop differs from these existing approaches and guidelines as it takes a comprehensive (socio-technical) engineering perspective: including all stakeholders (in addition to the defense organization, for example, regulators, subject-matter experts, and AI manufacturers). Additionally, the SOTEF loop focuses on the iterative nature of the human-AI system where continuous feedback, adaptation, and improvement throughout its lifecycle are essential. As such, it connects current approaches with each other and allows the functionality of the human-AI system to develop over time in a responsible way.

The SOTEF loop describes a process to (i) identify the ELSA to which the behavior of the human-AI system should adhere (including assigning responsibilities that apply during operation, (ii) ensure that the human-AI system can operate according to those

aspects, and (iii) enable stakeholders on different levels to regularly reflect and give feedback on the system's behavior and propose appropriate value-based adjustments. There is no single solution that achieves these three goals in all possible applications of AI technologies and this means that solutions are situation-dependent, that is, they are affected by the specific AI system deployed and the specific context of the governance, design, configuration, and operation. And because context, as well as values, change over time, the involvement and feedback from different stakeholders is needed during the complete lifecycle of the STS, from redesigns between iterations in order to realign to such new values to human support capabilities in order to intervene during operation when misalignment occurs. Instead of aiming for a one-size-fits-all solution, the SOTEF loop introduces a set of methods to identify and operationalize the relevant ELSA (given the mission goals) in order to establish MHC of a specific AI system in a specific context. The applicability of each method should be carefully considered in each specific context and might range from setting rules to which the human-AI system should adhere, predefining the behavior of the AI system, learning from human-selected data, to using goal functions and augmented utilitarianism (Aliman & Kester, 2022) (see definitions in Appendix 2.A).

We believe that the SOTEF loop is especially useful when dealing with high-risk military AI applications. Risk can be defined as the likelihood that unintentional harm of any kind (e.g., social, psychological, physical, or technical) can be done, with high risk implying that is it very likely that such harm will be done in the context of operation. Therefore, high-risk AI can be considered as unintentional harm being highly likely, as a result of the context in which the AI system is applied, the capabilities of the AI system, and/or the way in which it is applied (e.g., the human-AI interactions that take place). This makes MHC over AI systems in such contexts highly relevant, as the behavior that results from assessing situations and weighing values will have a major impact. Our assumption is that the higher the risk of the human-AI system application, the more extensive these moral considerations need to be for that risk to be mitigated. As a result, higher (ethical) demands are placed on the behavior of the human-AI system. If the design process results in a requirement that the AI system must base its behavior on moral values, then the AI system's implemented internal processes need to explicitly incorporate these values.

THE SOTEF LOOP

Santoni de Sio and Van den Hoven (2018) identified tracking and tracing requirements for meaningful control of autonomous AI systems, being: (i) responsiveness to the environment and moral considerations of the humans designing and deploying the AI systems, and (ii) providing the possibility to trace back the outcomes to a human during the design and operation process. We propose the SOTEF methodology as a way to explicate and embed these requirements in the design process for a human-AI system in a specific context. The SOTEF methodology operationalizes these requirements at

different control levels in four feedback loops: governance, design, development, and operation. The SOTEF methodology aims to structure the process for achieving MHC in an iterative fashion and offers validated methods for operationalization. Furthermore, the SOTEF methodology recognizes that this process and the methods used will differ per application, as each will be unique with respect to ELSA.

The SOTEF methodology prescribes how to set up the governance, design, development, and operation of a human-AI system. Each of these topics forms four distinct feedback loops at various timescales required in iteratively constructing and improving a human-AI system to behave according to ELSA. These intertwined feedback loops are based, respectively, on standardization efforts (Zielke, 2020), design processes such as the Double Diamond (Kunneman et al., 2022), system engineering processes such as the V-Model (Clark, 2009), and human-machine teaming interaction principles (Van der Waa et al., 2020). However, these processes are not explicitly intertwined with governance (Coeckelbergh, 2019). For this reason, the SOTEF loop includes a governance feedback loop to include regulation, policy, and laws in the construction and maintenance of human-AI systems.

A single method toward responsible military AI does not exist, as the possible applications of human-AI systems differ too greatly in terms of their goals, required tasks and capabilities, and ethical, legal, and societal context. A context that varies per society and the experiences that society acquired, and will continue to change over time as more experiences are gathered (Winston & Edelbach, 2013). This requires the development of multiple methods that can be applied in an iterative fashion for the human-AI system to adapt to a changing context. The feedback loops of the SOTEF methodology thus require varying methods, as the application and context demand. Methods that need to be developed and evaluated based on potential applications of human-AI systems and their ethical, societal, and legal contexts.

The feedback loops of governance, design, development, and operation are visualized in Figure 2.1. All are depicted as persisting feedback loops on various timescales. Each larger feedback loop governs the feedback loops on its lower timescale, while iteratively improving itself as it obtains feedback from those smaller loops. For example, the governance loop dictates the design process (e.g., NATO Principles signifying what should be considered during design), whereas the design process dictates what should be developed (e.g., how humans and AI systems should be interacting) and how the human-AI system should operate (e.g., as supervisory control). In turn, the design loop ideates on novel applications of human-AI systems that influence the governance loop. Below, we describe each loop and state its current challenges.

The Governance Feedback Loop

The governance loop dictates the principles and regulations for military applications of AI to provide the necessary ethical, legal, and societal context in which human-AI systems need to be designed, developed, and operated. This can include applying existing laws regulations, policies, and permissible opportunities for the military application of AI (cf. case law), or developing new ones. Examples are the international laws, regulations, standards, and guidelines established in the international human rights law,

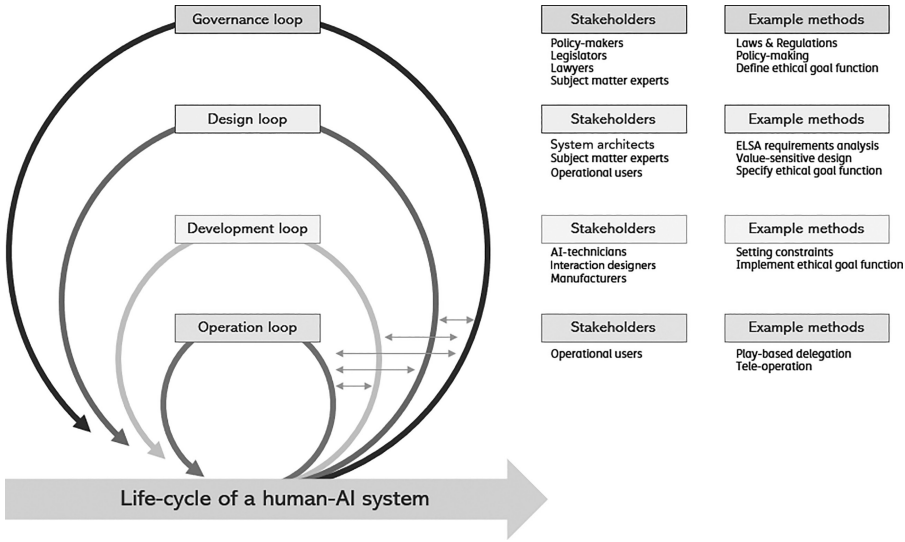

FIGURE 2.1 An illustration of the Socio-Technical Feedback (SOTEF) loop.

International humanitarian law, the EU AI Act, the high-level expert group on AI, ISO/IEC, and the NATO principles of responsible use of AI. The timescale of the governance loop is the largest of the four, as it attempts to be the most encompassing. It will provide the context that defines the inner, smaller loops. In turn, the smaller loops will provide critical reflection on whether the governance is sufficient or lacking in any way. Many efforts currently focus on the governance loop through methods such as committees and debates. However, these are often one-time exercises that lack the required iterative nature to match the progress and societal changes that occur over longer periods of time. Furthermore, such exercises only incorporate the feedback of specific AI applications in an ad-hoc fashion depending on societal events. Hence, the governance loop has various proven methods in place, but these methods need to be applied in a structured and iterative fashion.

The Design Loop

The design loop within the SOTEF methodology consists of an interdisciplinary design process that starts with determining the context, problem space, and the envisioned application of AI. Within this loop, the human-AI system should be defined, including a specification of the respective roles, tasks, goals, competencies, responsibilities, functions, and interactions of the humans and AI systems. The design loop should conclude with specified requirements on the human-AI system in the specific context that is considered, and in particular explicitly state the involved ELSA and how these should be addressed or implemented during development. Throughout this phase, the application of relevant laws, and the identification and specification of moral values and societal issues should come into play (Van de Poel, 2009). Numerous methods exist to support specific steps: from methods to involve potential or future users, for example,

Participatory Design (Schuler & Namioka, 1993; Ten Holter, 2022), to methods to integrate values in design and engineering, for example, Value Sensitive Design (Friedman & Kahn, 2003; Friedman & Hendry, 2019) or Rapid Ethical Deliberation (Aliman & Kester, 2022; Steen et al., 2021).

There are two main challenges in the design loop. The first challenge is the appropriate selection of the stakeholder(s) that need to be involved in selecting and carrying out subsequent design methods. Part of this selection involves determining the distribution of responsibility for the design across various stakeholders. The second challenge is the translation and specification of the identified values, ethics, and laws into concrete requirements that will guide the development process. The focus here is on specifying the desired functioning and behavior of the human-AI system. For example, what behaviors adhere to principles such as transparency and human agency, and values such as safety and integrity? What is required from the human-AI system to be able to show this behavior? For example, an AI system might be required to assess the safety of a given situation in order to warn a human operator when it encounters a dangerous situation. In order to be able to do so, it needs to be specified in the design loop what *safety* entails and how it can be assessed by the AI system. The design loop requires interaction with governance stakeholders to assess whether the system design warrants operation within existing laws and regulations. They are involved as a stakeholder in the design of the high-level capabilities and in the preparation of software specifications in order to indicate what developers must strictly adhere to and where they are allowed some flexibility in the implementation. This includes the design of processes for developers to reflect and communicate when the boundaries that have been set are in conflict with the given requirements (e.g., because of technical limitations). Thus a necessary discussion in any design loop is on who to involve with what responsibility to derive requirements from identified relevant moral values, ethics, and laws given the application that is considered. Methods are required that can facilitate this, on top of (existing) methods that shape the more general design process.

The Development Loop

The development loop should translate an agreed-upon design into a human-AI system while adhering to set translation restrictions to prevent diverging from the principles underlying the design. This translation effort should encompass all aspects of creating and maintaining a human-AI system. This loop includes all technological development efforts. In addition, it includes developing an organization to enable and support the designed human-AI system. Finally, it encompasses the validation and verification of the developed human-AI system as a whole. For example, a technical effort might include the development of a formal model that facilitates morally correct behavior for the AI system. Similarly, a new technical education and training regime can be developed that will impact human behavior. Finally, the resulting human-AI system – with trained humans and an AI system with a moral model – should be verified to assess whether it adheres to the specifications and validated whether it behaves as intended.

Two main challenges are part of the development loop: (i) the translation of design specifications into an implementation that adheres to the underlying ethical and legal aspects on which these specifications were based; and (ii) the reliable validation and

verification of such an implementation in light of these specifications. Methods are required to address these challenges, as it is essential to prevent any design choices from being made in the development loop (e.g., a developer making decisions on how a moral value should impact the human-AI system's behavior). If this occurs, the development loop would dictate (parts of) the human-AI system's behavior in an ill-supported manner.

The Operation Loop

Finally, the operation loop is the most inner feedback loop in the SOTEF methodology. This is where the actual human-AI interaction takes place, for example, to prepare or carry out a mission task. Here, the human-AI system is being instantiated, that is, it is decided when and how the human and AI-agents act (according to the predefined policies, rules, and requirements of the other loops) and where their behaviors can be observed. If the decision is not to deploy the AI, a reflection is required on the human-AI system's design and implementation. If the decision is in favor of deployment, the human-AI system must first be configured to tailor it to the foreseen deployment. This can include, for example, setting constraints for operation, providing specific training exercises, or specifying the task-specific goals of the system. There is a time gap between the development and operation of a human-AI system. A gap that is currently often neglected, as the discussion is about either how to design responsibly or how a human-AI system should operate. This ignores the fact that there is often time – and a necessity – to tailor the human-AI system to a particular deployment in a specific context. At times the moment for configuration is clear, for instance in the case of configuring autonomous AI systems just before a mission. At other times, this is much more diffuse, for instance in a classification algorithm running on always-on sensor systems. At those times, defining what constitutes configuration should be defined as part of the design process. Currently, methods are lacking to support this configuration component in the operational loop, so effort should be put into developing them.

The debate on MHC often focuses on either the governance, design, or operation loop. However, in practice, any control is likely to be a mix of control methods from governance, design, and operation. In the end, however, the final control mechanism resides in the operation. However, in some applications of AI, operational control is limited because *direct* human intervention in an AI agent's behavior is limited or even impossible. For instance, when communication is difficult or decisions need to be made in a short amount of time. However, within the SOTEF methodology, the operation loop is defined more broadly than mere direct human intervention. It also encompasses less direct interventions through reviews and feedback, which can be provided before, during, or after a specific instance of use (e.g., a specific military operation). Such methods (e.g., run time verification) can provide feedback to the other loops, potentially triggering a new design, development, or even new governance.

The goal of the SOTEF methodology is to ensure that the entire human-AI system behaves in a morally acceptable manner and abides by relevant governance while effectively achieving the set goals. The methodology takes a high-level perspective, giving opportunities to develop new control mechanisms where governance, design, and

development control mechanisms are intertwined with operational control. This is paramount for the responsible use of AI technologies. It broadens the discussion about how humans can control an AI system during operation, toward the discussion of how humans can control the behavior of a human-AI system in a combination of governance, design, development, and operational control.

Adaptation through Iteration

These four loops involve different human actors and act on their own timescales and should continue throughout the entire life cycle of the human-AI system. This means that the design of the system could always be reconsidered, a developed human-AI system should be open for change, and even during operations, the human-AI system should be able to adapt when needed. For example, humans might decide to not use the AI system during a specific operation as the context changes. To make this decision they might receive explanations from the AI system that convey the risks of using it, thus giving the AI system a role in the decision. A role that is defined in the design and development loop as the explanations are designed and implemented.

The SOTEF methodology recognizes that such complex and intertwined control mechanisms require iterations and places them in an overarching process during the lifetime of a human-AI system. These feedback loops are required as no matter the used methods, it cannot be guaranteed that a human-AI system always behaves responsibly according to ELSA in every possible situation. The SOTEF methodology addresses this by connecting governance, design, development, and operation activities over iterations to ensure the best possible human-AI system that is improved as experience is gathered – from sandbox environments to real operations.

The feedback loops allow for adaptation to changing circumstances, new insights, and changing values. Without these, the resulting human-AI system would be static in an ever-changing context which would eventually result in its failure to comply with the then-current moral values. As such, the SOTEF methodology recognizes not only that feedback occurs within each loop but also across loops, in both a top-down and bottom-up fashion. For example, the governance loop can change as new laws and regulations are implemented that set new requirements for human-AI systems. Similarly, these new laws and regulations can arise due to gained experience by already deployed human-AI systems as they pass through multiple operation loops.

The Involved Stakeholders, Their Responsibility and Accountability

Following the SOTEF methodology requires the involvement of many stakeholder groups. These include, but are not limited to, lawyers, subject matter experts, policy-makers, legislators, operational users, technicians, AI experts, manufacturers, and designers. The SOTEF methodology does not dictate that generic responsibilities should be assigned to each stakeholder group. Rather, it dictates that for each of the four

feedback loops methods should be available and responsibility should be assigned based on the selected methods and who should be involved in them. This selection is expected to occur according to governance and during the design, where governance would likely dictate relevant stakeholder groups, and in the design of a human-AI system specific representatives are selected.

The methods in the SOTEF methodology should be developed and evaluated on their contribution to the responsible use of AI for particular application domains. Such methods will dictate the participation and role of stakeholder representatives from which responsibility and accountability can be derived. It should be noted that there are dependencies between the loops and that there needs to be shared awareness of the outcomes across the loops (e.g., to assure that performance at the operations addresses the regulations of the governance). In addition, the overarching responsibility of developing and evaluating such methods lies with those involved in developing the SOTEF methodology further That is, those involved in the development of a method are *responsible* for communicating its strengths and limitations, while those who choose to apply a method are *accountable* for the implications of this method in the context of the application.

BOX 2.2 ILLUSTRATION OF SOTEF LOOP USING THE RAPID DEFENSE SCENARIO

SOTEF IN THE SCENARIO OF RAPID DEFENSE

Within the Rapid defense scenario, the iterative, transdisciplinary, value-sensitive approach of the SOTEF methodology can be illustrated as follows. Before the *RivalReveal* (RR) system is put into operation, weapon reviews of RR are performed (*governance*) using existing regulations. This informs the *design* of interaction and autonomy, for example, by stating requirements for operator interfaces and implementing automatic failure mode responses. Furthermore, training programs are developed for users to *configure* and *operate* the system. In this phase, collaboration between the various stakeholders is essential to manage the interdependencies between the cycles, for example, governance dictates design choices, and design possibilities inform compliance. After the initial system is evaluated, the system is taken into use. As the SOTEF methodology is a lifecycle approach, it does not end there. Compliance with ethical guidelines such as appropriate levels of judgment and care are continuously monitored against their effectiveness. On all longer timescales, this can lead to changing the regulatory framework, for example, changing the ethical principles, or sharpening them to be more precise. At the levels of design and operation, incidents and practical experiences with RR are continuously monitored and are assessed in relation to public values. These insights could be used at all levels: to train operators, sharpen system design, and even to reconsider ethical guidelines based on the way they play out in practice. A functioning SOTEF methodology does not arise naturally: it requires careful consideration of all stakeholders, providing them with the right information at the right moment and empowering them to act.

METHODS AND FUNCTIONS IN THE SOTEF LOOP

Responsible military AI must be achieved by empowering humans to align the behavior of the STS with human values. The SOTEF loop facilitates this process by offering various control methods and functions for identifying, implementing, monitoring, and adjusting relevant values and goals in a STS. Methods are concrete processes and procedures that can be used to implement part of a control loop. For example, utility elicitation and value-sensitive design are methods to identify relevant moral values within a domain, which is one of the goals in the Design loop. Functions are the prescribed capabilities of the STS that should be implemented in a specific component (e.g., AI system, human, environment) of the STS. For example, explainability of system behavior might be a function that is required for the Governance loop, as those who govern need to understand the relationship between the applicable regulation and the behavior of the system in the operational context for which they might be held accountable. Team Design Patterns can be used to explore the allocation of functions in the STS (Van der Waa et al., 2020).

Table 2.1 lists methods and functions that can be used for the instantiation of the SOTEF loop methodology. Note that methods and functions can be applied in combination and might entail interdependencies. For example, an ethical goal function (method) can prescribe the selection of training data (function). As Table 2.1 shows, the methods and functions that are being used in a SOTEF-loop may take a variety of forms, differing with respect to:

- The **component** of the STS for which the method is implemented (e.g., the AI system, the human, the environment, the interaction, etc.)
- The **response time** between the act of controlling and the moral behavior (outcome of moral decision-making) of the STS that is being controlled. *Long* means more than one day, *medium* means between ten seconds and one day, *immediate* means less than ten seconds
- The **human actor** that executes control over the STS
- The **feedback mechanism** by which the STS's behavior is monitored and used as input to further control the STS (e.g., to realign the STS with relevant moral values)

The repertoire of methods and functions that can be applied to instantiate the feedback loops will evolve over time. The sections above already mentioned value sensitive design (Friedman & Kahn, 2003; Friedman & Hendry, 2019), rapid ethical deliberation (Steen et al., 2021), participatory design (Bratteteig & Verne, 2018), and moral programming (Aliman & Kester, 2022). Table 2.1 shows other relevant methods and functions as a further illustration of the repertoire of methods. Current research of the ELSA labs in the Netherlands[11] will extend this list and provide a comprehensive state-of-the-art overview.

TABLE 2.1 Methods and functions to instantiate (part of) the feedback loops of the SOTEF-methodology

NAME	*LOOP*	*COMPONENT*	*RESPONSE TIME*	*HUMAN ACTOR*	*FEEDBACK MECHANISM*	*EXAMPLE*
Restricting use context	Governance	Legal Context	Long	Legislators and policymakers	Law enforcement	Prohibition stating that AI system must not be used in urban environments
Value identification	Governance/ Design	Human, Ethical context	Long	Various stakeholders	Value deliberation and validation	Identifying ethical considerations for AI healthcare applications (Char et al., 2020)
Requirement analysis	Design	Human, AI, inter-action	Long	Military authorities, Human-AI interaction experts	Requirement validation	Scenario-based requirements engineering (Sutcliffe, 2003)
Algorithm auditing	Governance Design Development	AI	Long	Various stakeholders	Explainable AI	Risk rating, surrogate explanations, post-processing, etc. (Koshiyama et al., 2022)
Defining ethical decision framework	Design	Human, AI, inter-action	Long	Military authorities, Human-AI interaction experts	Decision quality validation	Allocating ethical decision-making in a human-AI team (van der Waa, 2020)
Shaping infrastructure	Design	Environmental context	Long	Military planners	Incident management system	Placing a fence around the AI's workplace
Selecting training data	Development	AI	Long	AI engineers	Explainable AI	Engineers compose image datasets of representative "hostile vehicles"
Ethical goal function	Governance Design Development	AI	Long	Various stakeholders	Explainable morality	Value and harm model in an autonomous car (Reed et al., 2021)

(*Continued*)

TABLE 2.1 (*Continued*) Methods and functions to instantiate (part of) the feedback loops of the SOTEF-methodology

NAME	*LOOP*	*COMPONENT*	*RESPONSE TIME*	*HUMAN ACTOR*	*FEEDBACK MECHANISM*	*EXAMPLE*
Norm engineering	Development	AI	Long	AI engineers	Explainable AI	Privacy-enhancing technologies (e.g., Liu et al., 2021)
Human task training	Operation	Human	Long	Military trainers, doctrine developers	Incident management system	Training a soldier to work with a particular AI system
Human resilience training	Operation	Human	Long	Military trainers	Simulation-based training	Appraisal training to decrease the effects of traumatic experiences (Beer et al., 2020)
Play-based delegation	Operation	Inter-action	Medium	Human teammate	Progress appraisal	Human calls predefined play for doing an area surveillance during a mission (Miller & Parasuraman, 2007; van Diggelen et al., 2021)
Collaborative planning	Operation	Inter-action	Medium	Human teammate	Progress appraisal, AI-assisted feedback	Human and AI formulate mission plan together
Tele-presence	Operation	Inter-action	Immediate	Tele-operator	Visual, sound & other senses	AI autonomously performs surveillance, but is taken over by the human in unexpected situations
Adaptive automation	Operation	Inter-action	Immediate	Operator	Adjustable work agreements, explaining displays	Attuning the level of automation to the momentary situation and operator workload (De Tjerk et al., 2010)

Note: Response time is a relative concept; the context and momentary risk level of the (planned) operation determine its actual value.

Note that each of these methods and functions can be implemented in the SOTEF loop. The outcome of a specific method does not guarantee ethically aligned behavior of the STS over time. Under the assumption that a combination of methods will lead to better ethically aligned behavior, verification and validation should entail the comprehensive sum of these outcomes, that is, the results of all loop levels. As the human, AI, and environment change over time due their "inner" feedback loops *and* their adaptations to each other, regular reviews and adjustments should be made to the STS during its complete life cycle. This process should be directed by those who have an overview of – and insight in – the STS in the context in which it is deployed. Since different stakeholders are involved in each loop, the review and adjustment process has to take place in interaction with each other.

DISCUSSION AND CONCLUSION

In this chapter, we provided an overview of the SOTEF methodology: a comprehensive, iterative STS-engineering approach that distinguishes a governance, design, development, and operation loop for responsible military AI life cycles. The implementation of these feedback loops will be done within a specific context for a specific set of objectives, affecting (1) the scope and types of moral considerations and (2) the choice and modes of AI applications. Table 2.1 presents a set of methods and functions that can be applied to instantiate the loops (with their distinctive features and some examples). Such instantiations involve the combination of the most appropriate methods and functions to establish the desired situated value alignment and MHC. An illustrative scenario exemplified the proposed value-alignment process for MHC (i.e., how the SOTEF implementation can be achieved). There will be some challenges to fully implement the SOTEF methodology. We will briefly discuss these below.

One challenge is to select relevant stakeholders at an early stage and to provide them with the needed resources. The involvement of stakeholders is key in the SOTEF loop, and should already be arranged at the start of an exploration or a design of AI functionality for military operations. However, stakeholder involvement (e.g., legal experts, legislators, ethicists, military users, system engineers, AI developers, and NGOs) has its challenges in all feedback loops. For example, stakeholders might want to protect themselves from co-optation by other stakeholders, safeguarding freedom of speech and maintaining independence (confidentiality). Another practical example concerns resources. Time and money may constrain relevant stakeholders to participate in value dialogues (Krabbenborg, 2020).

Another challenge is that the engagement of stakeholders also raises questions about communication and empowerment. Although the SOTEF loop relies on ideals of willingness to cooperate, openness, and harmony, it is known that these ideals are rarely realized in practice (Blok, 2014). The SOTEF methodology will provide the arguments and tools to establish the required engagement, referring to the applicable standards, methods, and functions. Furthermore, we aim to build up and share experiences on "who to involve how" (e.g., the implementation of the stakeholder roles and

involvement of representatives of "unaccustomed" stakeholder groups such as citizens in a mission area), and how different values can be conceptualized, expressed, reported, and balanced.

The third challenge is that humans may find it hard to acknowledge, explicate and verbalize their values, because their primary assessment of right and wrong is often *implicit*, based more on emotional responses and less on rational (conscious) considerations (Haidt, 2001; Van Diggelen, Metcalfe, Van den Bosch, Neerincx, & Kerstholt, 2023). Furthermore, people's moral assessments are not unitary but *multi-dimensional* and *context-dependent* (i.e., related to the specific situation, work, and social roles; cf., Hannah, Thompson & Herbst, 2020; Aliman & Kester, 2022). The provision of scenarios, vignettes, and simulations in a virtual reality environment might help to systematically reflect on the moral aspects at stake, making the implicit explicit (cf., Parsons, 2015).

In conclusion, the SOTEF loop methodology comprises the assessment of a specific human-AI system operating in a specific context through an iterative, transdisciplinary, and multistakeholder approach. Although military AI creates new challenges and concerns for moral decision-making, it can also provide part of the solution. The use of military AI forces us to think about what values are at stake and how we want to ensure these values. SOTEF supports making ELSA of human-AI system deployment explicit, comparable, and auditable. It provides a way to better explicate attribution of responsibility and accountability; as such it is a way forward to operationalize MHC of military AI-based systems. It challenges stakeholders to make explicit and validate their goals and moral values, for the specific context the human-AI system is to operate in. Currently, we are operationalizing this methodology for realistic use cases, in order to refine and test the applicability of various methods and functions.

ACKNOWLEDGMENTS

We would like to express our gratitude to Dr. A.W. Bronkhorst and J.A.P. Smallegange for their invaluable contributions to this work. Their insightful comments and feedback during discussions helped us to refine our ideas and approaches. We also thank them for taking the time to review the chapter several times, and for providing constructive criticism that has helped us to clarify our arguments and conclusions.

APPENDIX 2.A: GLOSSARY

The table below provides working definitions of core concepts in this chapter. The TAILOR Handbook of Trustworthy AI provides a more generic overview of relevant definitions of trustworthy AI in the form of a publicly accessible Wiki: http://tailor.isti.cnr.it/handbookTAI/TAILOR.html.

CONCEPT	*WORKING DEFINITION*
Morality and ethics	Both morality and ethics pertain what is right ("good") and wrong ("bad"). The word morality is more used in relation to the personal normative aspects, whereas ethics more in relation to the normative standards within a certain community or social setting.
High-risk	The likelihood that unintentional socio-psycho-techno-physical perceived serious harm can be done.
Moral model	A formal model that represents what is right and what is wrong (e.g., in terms of action's benefits and harms) and, as such, univocally governs the behavior of the socio-technical system (STS). An AI agent's moral model is a formal model of how it should behave such that it contributes to a morally acceptable behavior of the STS as a whole.
Socio-technical system	A holistic perspective of a system containing an interconnection between humans (society as a whole) and (AI-based) technologies, including both social and technical aspects.
Socio-technical feedback loop	The human-centered methodology that addresses the context of all stakeholders of an AI application comprehensively, and prescribes a life-cycle enduring review and refinement process to enhance the models, reasoning, and adaptations to changing circumstances.
Value-alignment	The continuous process including the identification of the moral values that are at stake, and how they are addressed in a military operation.
Human-AI system	All of the humans and AI agents combined that collaborate to achieve a shared goal during operation.
Methodology	A methodology is a set of methods employed by a discipline. In the context of this chapter, it is the discipline of arriving at a responsible application of AI in the military domain.
Moral value	Something held to be right/wrong or desirable/undesirable at a certain moment in time by a certain group of people. Moral values describe what people value in terms of what they believe is morally acceptable. Fundamental examples include honesty and respect. Pragmatic examples include being fair and respecting another's privacy.
Goal function	A model of what the AI application should pursue such that it can be used to steer the AI application's behavior. Examples include optimization functions (loss, reward, fitness, utility functions) and logic inference rules (drawing conclusions and rule resolution).
Augmented utilitarianism	A non-normative meta-ethical framework that builds upon the foundational principles of deontological ethics, consequentialist ethics, and virtue ethics and combines them in one framework. Augmented utilitarianism tries to capture a more nuanced and comprehensive understanding of human harm perception from the perspective of moral psychology, for example, "dyadic morality". AU functions as a scaffold to encode human ethical and legal conceptions in a machine-readable form (e.g., ethical goal functions).
Ethical goal function	A goal function that also models moral values and thus governs an AI application's behavior in terms of what should be pursued in terms of how those values are modeled. Examples include multiobjective functions, utility functions, or multicriteria optimization functions whose attributes approximate observable moral values, and inference engines whose inference rules incorporate deontic logic.

NOTES

1 https://www.stopkillerrobots.org/
2 IHL Treaties – Geneva Convention (III) on Prisoners of War, 1949 – Article 3 (icrc.org)
3 TER report HFM-RWS 322: meaningful human control (MHC) of artificial intelligence (AI)-based systems, https://scienceconnect.sto.nato.int/tap/activities/11639
4 STS can refer to either socio-technical or socio-technological system. They both relate to the interaction between social and technical elements in a human-AI system, but might emphasize slightly different aspects. The former is more often used to emphasize the need for a holistic approach focusing both on technical and human factors, while the latter is used to highlight the role of technology in shaping interactions, behaviors and outcomes of a human-AI system. We use the term socio-technical as this is the more established and commonly used term in scientific literature.
5 AIV/CAVV advice 2021 and cabinet response 2022: https://www.adviesraadinternationalevraagstukken.nl/documenten/publicaties/2021/12/03/autonome-wapensystemen
6 Guiding Principles affirmed by the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons System: https://www.ccdcoe.org/uploads/2020/02/UN-191213_CCW-MSP-Final-report-Annex-III_Guiding-Principles-affirmed-by-GGE.pdf
7 NATO Principles of Responsible Use: https://www.nato.int/docu/review/articles/2021/10/25/an-artificial-intelligence-strategy-for-nato/index.html
8 The TAILOR Handbook of Trustworthy AI (http://tailor.isti.cnr.it/handbook-TAI/TAILOR.html).
9 https://elsalabdefence.nl/
10 https://rri-tools.eu/
11 https://nlaic.com/en/category/building-blocks/human-centric-ai/elsa-labs-en/

REFERENCES

Aliman, N.-M. (2020). Hybrid cognitive-affective strategies for AI safety [Doctoral dissertation, Utrecht University]. https://doi.org/10.33540/203

Aliman, N. M., & Kester, L. (2022). Moral programming: Crafting a flexible heuristic moral meta-model for meaningful AI control in pluralistic societies. In *Moral design and technology* (pp. 494–503). Wageningen Academic Publishers.

Aliman, N.-M., Kester, L., Werkhoven, P., & Yampolskiy, R. (2019). Orthogonality-based disentanglement of responsibilities for ethical intelligent systems. *In International conference on artificial general intelligence (AGI)*, Shenzhen.

Amoroso, D., & Tamburrini, G. (2020). Autonomous weapons systems and meaningful human control: Ethical and legal issues. *Current Robotics Reports*, *1*(4), 187–194.

Beer, U. M., Neerincx, M. A., Morina, N., & Brinkman, W. P. (2020). Computer-based perspective broadening support for appraisal training: Acceptance and effects. *International Journal of Technology and Human Interaction (IJTHI)*, *16*(3), 86–108.

Blok, V. (2014). Look who's talking: Responsible innovation, the paradox of dialogue and the voice of the other in communication and negotiation processes. *Journal of Responsible Innovation*, *1*(2), 171–190.

Bratteteig, T., & Verne, G. (2018). Does AI make PD obsolete? Exploring challenges from artificial intelligence to participatory design. *In Proceedings of the 15th participatory design conference: Short papers, situated actions, workshops and tutorial-volume 2* (pp. 1–5).

Clark, J. O. (2009, March). System of systems engineering and family of systems engineering from a standards, V-model, and dual-V model perspective. In *2009 3rd annual IEEE systems conference* (pp. 381–387). IEEE.

Coeckelbergh, M. (2019). Artificial intelligence: Some ethical issues and regulatory challenges. *Technology and Regulation*, 2019, 31–34.

De Greef, T. E., Arciszewski, H.F.R., Neerincx, M. A. (2010). Adaptive automation based on an object-oriented task model: Implementation and evaluation in a realistic c2 environment. *Journal of Cognitive Engineering and Decision Making*, *4*(2), 152–182.

De Greef, T. E., Arciszewski, H.F.R., Neerincx, M. A. (2020). Identifying ethical considerations for machine learning healthcare applications. *The American Journal of Bioethics*, *20*(11), 7–17.

Dunnmon, J., Goodman, B., Kirechu, P., Smith, C., & Van Deusen, A. (2021). *Responsible AI guidelines in practice: Lessons learned from the DIU portfolio.* Washington, DC: Defense Innovation Unit.

Ekelhof, M. (2019). Moving beyond semantics on autonomous weapons: Meaningful human control in operation. *Global Policy*, *10*(3), 343–348.

Friedman, B., & Hendry, D. G. (2019). *Value sensitive design: Shaping technology with moral imagination*. Cambridge, MA: MIT Press.

Friedman, B., & Kahn, P. (2003). Human values, ethics, and design. In J. Jacko & A. Sears (Eds.), *The human-computer interaction handbook: Fundamentals, evolving technologies and emerging applications* (pp. 1177–1201). Mahwah, NJ: Lawrence Erlbaum Associates.

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, 108, 814–834.

Hannah, S. T., Thompson, R. L., & Herbst, K. C. (2020). Moral identity complexity: Situated morality within and across work and social roles. *Journal of Management*, *46*(5), 726–757.

Koshiyama, A., Kazim, E., & Treleaven, P. (2022). Algorithm auditing: Managing the legal, ethical, and technological risks of artificial intelligence, machine learning, and associated algorithms. *Computer, 55*(4), 40–50.

Krabbenborg, L. (2020). Deliberation on the risks of nanoscale materials: Learning from the partnership between environmental NGO EDF and chemical company DuPont. *Policy Studies*, *41*, 372–391

Kunneman, Y., Alves da Motta-Filho, M., & van der Waa, J. (2022). Data science for service design: An introductory overview of methods and opportunities. *The Design Journal*, *25*(2), 186–204.

Liu, X., Li, H., Xu, G., Chen, Z., Huang, X., & Lu, R. (2021). Privacy-enhanced federated learning against poisoning adversaries. *IEEE Transactions on Information Forensics and Security, 16*, 4574–4588.

Miller, C. A., & Parasuraman, R. (2007). Designing for flexible interaction between humans and automation: Delegation interfaces for supervisory control. *Human Factors*, *49*(1), 57–75.

Parsons, T. D. (2015). Virtual reality for enhanced ecological validity and experimental control in the clinical, affective and social neurosciences. *Frontiers in Human Neuroscience*, *9*, 660.

Peeters, M. M., van Diggelen, J., Van Den Bosch, K., Bronkhorst, A., Neerincx, M. A., Schraagen, J. M., & Raaijmakers, S. (2021). Hybrid collective intelligence in a human-AI society. *AI & Society*, *36*, 217–238.

Reed, N., Leiman, T., Palade, P., Martens, M., & Kester, L. (2021). Ethics of automated vehicles: Breaking traffic rules for road safety. *Ethics and Information Technology*, *23*(4), 777–789.

Santoni de Sio, F., & Van den Hoven, J. (2018). Meaningful human control over autonomous systems: A philosophical account. *Frontiers in Robotics and AI*, *5*, 15.

Scharre, P. (2018). *Army of none: Autonomous weapons and the future of war.* WW Norton & Company.

Schuler, D., & Namioka, A. (1993). *Participatory design: Principles and practices.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Steen, M., Neef, M., & Schaap, T. (2021). A method for rapid ethical deliberation in research and innovation projects. *International Journal of Technoethics (IJT)*, *12*(2), 72–85.

Stilgoe, J., Owen, R., & Macnaghten, P. (2013). Developing a framework for responsible innovation. *Research Policy*, *42*, 1568–1580

Sutcliffe, A. (2003). Scenario-based requirements engineering. In Proceedings 11th IEEE *international requirements engineering conference, 2003* (pp. 320–329). IEEE.

Ten Holter, C. (2022). Participatory design: lessons and directions for responsible research and innovation. *Journal of Responsible Innovation*, *9*(2), 275–290.

Van de Poel, I. (2009). Values in engineering design. In A. Meijers (Ed.), *Handbook of the philosophy of science. Volume 9: Philosophy of technology and engineering sciences* (pp. 973–1006). Elsevier.

Van den Hoven, J. (2013). Value sensitive design and responsible innovation. In R. Owen, J. Bessant, and M. Heintz (Eds.) *Responsible innovation* (pp. 75–83). (Chichester, UK: John Wiley & Sons, Ltd).

van der Waa, J., van Diggelen, J., Cavalcante Siebert, L., Neerincx, M., & Jonker, C. (2020). Allocation of moral decision-making in human-agent teams: A pattern approach. In *Engineering psychology and cognitive ergonomics. Cognition and design: 17th international conference, EPCE 2020, held as part of the 22nd HCI international conference, HCII 2020, Copenhagen, Denmark,* July 19–24, 2020*,* Proceedings, Part II 22 (pp. 203–220). Springer International Publishing.

van Diggelen, J., Barnhoorn, J., Post, R., Sijs, J., van der Stap, N., & van der Waa, J. (2021). Delegation in human- machine teaming: Progress, challenges and prospects. In *Intelligent human systems integration 2021: Proceedings of the 4th international conference on intelligent human systems integration (IHSI 2021): Integrating people and intelligent systems, February 22–24, 2021, Palermo, Italy* (pp. 10–16). Springer International Publishing.

van Diggelen, J., Metcalfe, J. S., Van den Bosch, K., Neerincx, M, & Kerstholt, J. (2023). Role of emotions in responsible military AI. *Ethics and Information Technology*, *25*, 17. https://doi.org/10.1007/s10676-023-09695-w

Van Veenstra, A. F., Van Zoonen, L., & Helberger, N. (2021). *ELSA Labs for human centric innovation in AI.* Netherlands AI Coalition.

Von Schomberg, R., & Hankins, J. (Eds.). (2019). *International handbook on responsible innovation: A global resource.* Edward Elgar.

Winston, M., & Edelbach, R. (2013). *Society, ethics, and technology.* Cengage Learning.

Zielke, T. (2020). Is artificial intelligence ready for standardization? In *Systems, software and services process improvement: 27th European conference, EuroSPI 2020, Düsseldorf, Germany, September 9–11, 2020, proceedings 27* (pp. 259–274). Springer International Publishing.